Author:
Rightmyer, Jake

Title:
Using and improving whole-cell models to investigate bacterial minimal genomes

Using and Improving Whole-Cell Models to Investigate Bacterial Minimal Genomes

Jake Rightmyer

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Master of Science by Research (MScR) in the Faculty of Life Sciences, School of Biological Sciences.

May 2021

Word count: 19,370

Abstract

We are approaching meaningful incorporation of *in silico* whole-cell modelling approaches and *in vivo* custom genome design. The whole-cell model of *Mycoplasma genitalium,* published in 2012, has permitted investigation into the organism's metabolism and genome that has been impossible *in vivo*. The ongoing development of a whole-cell model for *Escherichia coli*, a well-annotated model organism, has presented us with the opportunity to test *in silico* predictions from a whole-cell model *in vivo* for the first time.

In this thesis, I investigate *in silico* predictions made by both the *M. genitalium* and the *E. coli* whole-cell models. Through gene ontology term analysis of eight minimal gene sets found in the literature, I identified the biological functions that *M. genitalium* cells can dispense with *in silico* to produce dividing cells. My results clarify which biological functions should be included in any minimal gene set for *M. genitalium*. I investigated the results of 1214 single gene knockouts in the *E. coli* whole-cell model and compared these results with the assessment of essentiality in the Keio collection. At least 68.8% of knockouts agreed with the Keio collection over two generations, indicating the *E. coli* whole-cell model can predict gene essentiality with some accuracy. I also identify ways to develop the model to improve predictions. I prepared to use the no-SCAR system of genomic engineering in our laboratories to perform a proof-of-concept deletion in *E. coli* and enable *in vivo* investigation of *E. coli* whole-cell model predictions. I then worked with the Covert Lab, Stanford, to continue to develop the *E. coli* whole-cell model, curating data on the thiosulfate sulfurtransferase reaction and Lrp transcriptional regulation from the literature.

Despite interruptions caused by the Covid-19 pandemic, the work reported in this thesis represents progress towards the clarification and application of whole-cell models to investigate bacterial genomes.

Acknowledgements

I would like to thank my supervisors, Claire Grierson and Lucia Marucci, for their guidance, resilience, and belief in me to achieve the completion of this MScR degree under ever-changing circumstances. I have loved my time as a member of their research group and will cherish the experiences they have made possible.

This project would have been impossible to complete without the mentorship of Joshua Rees-Garbutt whose hands-on teaching concerning everything from coding to lab techniques, and a crash course in whole-cell models allowed me to develop confidence, and I daresay, competence in this field. I am extremely grateful for the time invested in me, not to mention the opportunity to be published as his second author on two occasions.

I thank Sophie Landon for all her invaluable assistance with coding and to get the *E. coli* WCM up and running. When every effort of mine ended in a different error message, I could always rely on a step-by-step tutorial to set me straight.

I thank Ioana Gherman for kindly running simulations for me when I lacked the time to do so myself and I wish her the best for the rest of her PhD.

I would like to thank the Covert Lab at Stanford University for their enthusiasm for collaboration with our group. In particular, Travis Horst and Gwanggu Sun for their involved efforts to educate myself and Sophie on the inner workings of their exciting model.

I would finally like to thank my family, friends, and housemates for their continued love and support in these unprecedented and frankly, quite strange times.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: Jake Rightmyer DATE: 28th May 2021

Contents

List of Figures

List of Tables

Chapter 1 – Introduction

1.1 - Statement of Collaboration

Elements of work presented in this thesis have been published in Current Opinion in Systems Biology and are under review for potential publication in ACS Synthetic Biology. Any work included in this thesis is solely my contribution to these publications. I indicate where I use scripts or data produced by others.

- Rees-Garbutt, J., Rightmyer, J., Karr, J. R., Grierson, C., & Marucci, L. (2020a). Furthering genome design using models and algorithms. *Current Opinion in Systems Biology*, *24*, 120–126.
    - Second author

- Rees-Garbutt, J., Rightmyer, J., Chalkley, O., Marucci, L., & Grierson, C. (2020b). Testing theoretical minimal genomes using whole-cell models. *BioRxiv*, 2020.03.26.010363.
    - Second author

1.2 - Current genome engineering

Genetic engineering is the deliberate modification of the characteristics of an organism by manipulating its genetic material. Genome engineering is genetic engineering applied to genomes (Carr & Church, 2009). The process of genome design, whereby the desired genetic modifications are selected and tested subsequently *in vivo* (Haimovich *et al.,* 2015) is time-consuming and expensive due to the limitations of current techniques and the inability to backtrack (Rees-Garbutt *et al.,* 2020c). Implementing processes to test the design of genomes *in silico* before conducting *in vivo* editing would incur fewer time and monetary costs than *in vivo* research alone and is therefore desirable.

The publication of whole-cell models (WCMs) (Karr *et al.,* 2012; Macklin *et al.,* 2020) and the development of clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) technology (Jinek *et al.,* 2012; Jiang *et al.,* 2013), when coupled with genome design algorithms (Rees-Garbutt *et al.,* 2020c) present us with the opportunity to combine *in silico* design and *in vivo* editing and begin to design genomes using these components. Due to the difficulty of working with *Mycoplasma genitalium* (*M. genitalium*) in the laboratory (Hutchison *et al.,* 2016), the

development of the *Escherichia coli* (*E. coli*) WCM will allow *in vivo* investigation of WCM predictions as *E. coli* is easily handled in the laboratory (Blount, 2015).

Genome minimisation has been used as a proof-of-concept of genome engineering and design as it provides a simple functional assay whether the cell replicates or not (Rancati *et al.,* 2018; Rees-Garbutt *et al.,* 2020c). Minimal genomes, cells containing only genes essential for survival (Glass *et al.,* 2017), have been proposed for *M. genitalium* and *E. coli* (Hutchison *et al.,* 1999; Yang *et al.,* 2019). As the simplest iteration of genome design, it may be possible to combine the technologies of WCMs, CRISPR-Cas9, and genome design algorithms to accelerate the design and production of organisms with minimal genomes.

1.3 - Genome-scale metabolic models and whole-cell models

Genome-scale metabolic models (GEMs) are *in silico* representations of a cell's metabolism and are formed from networks of metabolic reactions that contain all metabolites and genes encoding enzymes involved in metabolism (Price *et al.,* 2004). The overarching aims of this research are multifaceted, with goals including better understanding of cellular processes by comparing experimental data to model predictions, guided and systematic manipulation of organisms at the genome level, and elucidation of uncharacterised processes and genes (McCloskey *et al.,* 2013). GEMs are built by integrating biochemical metabolic pathways with annotated genome sequences, using information on enzymes, genes, reactions, pathways, and metabolites. Reconstructions of GEMs, using the elements mentioned above, were initially performed manually (Oberhardt *et al*., 2009). However, now there are tools that allow for semi-automatic assembly of reconstructions, including Pathway Tools for building models from the online database, EcoCyc. These tools include AutoKEGGRec, AuReMe, CarveMe, MetaDraft, and RAVEN amongst many others (Karlsen *et al.,* 2018; Aite *et al.,* 2018; Machado *et al.,* 2018; Hanemaaijer *et al.,* 2017; Wang *et* al., 2018). These are comprehensively reviewed by Mendoza *et al.* (2019). GEMs are invaluable for cell biologists, giving a means of visualising how metabolism in a cell functions and providing predictive capabilities for metabolic network responses to varying conditions such as altered substrate concentration, gene knockouts, and synthetic lethality (McCloskey *et al.,* 2013). It is possible to simulate the effects of manipulating metabolism within the *in silico* model, for example via Flux Balance Analysis (FBA) (Varma & Palsson, 1994). FBA permits the customised direction of the proportion of flux through different metabolic pathways, and as such provides a solution space where one can visualise the effects of scaling up or scaling down certain metabolic reactions on processes such as metabolite concentrations and cell growth (Orth *et al*., 2010). In essence, gene modifications act as the input in this system, with the altered flux distribution as the output (Landon *et al.,* 2019).

Until researchers began attempts to construct a WCM, GEMs served as the most complete way to computationally visualise cellular processes, permitting *in silico* research concerning metabolite optimisation, gene essentiality, and wild-type cell behaviour in 78 bacterial species (Oberhardt *et al.,* 2009; UCSD, 2018). GEMs have also inspired early development of more detailed extended models that take transcription processes into account, as well as macromolecular expression models (ME-models), which integrate macromolecular synthesis reactions alongside metabolic reactions, producing models with even greater fidelity to living cells (Ma *et al*., 2017). The above types of model are of great importance, with submodels utilising FBA, for example, providing some of the building blocks for the first WCM (Karr *et al.,* 2012). However, as this thesis is about using WCMs, I will mostly focus on those.

Since the creation of the first GEM for *Haemophilus influenzae* (Edwards & Palsson, 1999), the field of genome modelling has seen vast progression. WCMs are computational models which utilise mathematical integration of either cellular parameter values (Macklin *et al.,* 2020) or submodels (Karr *et al.,* 2012) to simulate the life cycle of an entire cell. WCMs are also capable of modelling individual molecules and interactions, including the function of all gene products in the cell (Landon *et al.,* 2019). Their creation has been revolutionary for *in silico* cellular research, facilitating extensive comparison of the literature against itself (Macklin *et al.,* 2020) as well as allowing research to be carried out which is either too expensive or simply unfeasible *in vivo* (Waltemath & Wolkenhauer, 2016). Two WCMs have been published: *M. genitalium* (Karr *et al.,* 2012) and *E. coli* (Macklin *et al.,* 2020) (Figure 1.1). WCMs for *Mycoplasma pneumoniae* (*M. pneumoniae*), H1 human embryonic stem cell, and an archetypal bacterium remain in development (Whole-Cell Modeling; URL: https://www.wholecell.org/models/). Of the extant WCMs, *M. genitalium* has 100% of well-annotated genes modelled (401 of 525 genes) (Karr *et al.,* 2012), while *E. coli* has 43% of well-annotated genes modelled (1214 of 4401 genes) (Macklin *et al.,* 2020). Remaining genes consist of genes of unknown function and, in the case of the *E. coli* WCM, genes yet to be implemented at the time of writing. Despite not having all well-annotated genes implemented, the *E. coli* WCM represents a vital progression in WCM creation, comprising of tenfold more parameter values than the *M. genitalium* WCM*,* and completing simulations in a fraction of the time while including 50 times the number of molecules as its predecessor (Macklin *et al.,* 2020). The *M. genitalium* WCM has also permitted the discovery of novel *in silico* minimal genomes, including the most reduced genome to date (Rees-Garbutt *et al.,* 2020c), a feat which could be replicated, and experimentally verified, when the *E. coli* WCM is complete. Custom genome design of a genome-reduced industrially useful organism may soon become a reality.

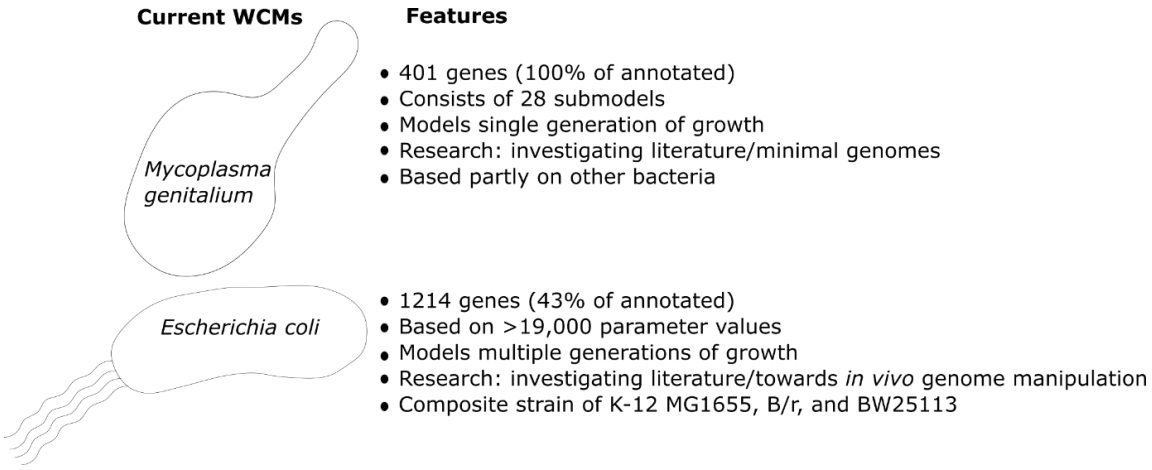

Figure 1.1. The two extant whole-cell models (WCMs) and some of their defining features (adapted from Rees-Garbutt *et al.,* 2020a).

To achieve genome-driven engineering of cells, it is necessary to marry metabolic and genomic engineering to control a large proportion of cellular processes, enacting change in both specific metabolite yields and in the genome of the cell (Landon *et al.,* 2019). As WCMs represent an *in silico* medium for modelling metabolism as well as cellular behaviour in response to genetic alterations, they provide a unique solution to streamline the production of genome-engineered cells. The *E. coli* WCM will permit the first attempts to engineer genomes *in vivo* with strategy informed by *in silico* WCM simulations, reducing the trial-and-error nature of the work, saving researchers time and money.

1.4 - Using whole-cell models

To understand the research presented in this thesis, it is necessary to provide some background information about WCMs including how they function and how they can be interacted with to produce useful data. The *M. genitalium* WCM functions by the integration of 28 submodels that are responsible for cellular processes such as metabolism and transcription and are mathematically modelled (Karr *et al.,* 2012). The model is run by using an algorithm that takes randomly initialised cell variables, which represent the complete configuration of the cell, and allocate them among the 28 submodels. These submodels then operate for one second before the cell variables are updated repeatedly until the cell divides or the maximum simulation time (13.89 hours) is reached (Karr *et al.,* 2012). The *E. coli* WCM is constructed from unified datasets defining cell parameters such as the numbers and types of RNA and protein molecules, upon which 19 algorithms comprising over 10,000 mathematical equations are run to simulate a cell growing and dividing (Macklin *et al.,* 2020). Macklin *et al.* (2020) analogise this process to the numerical integration of ordinary differential equations (ODEs), where the cellular states are analogous to the ODE state variables and the cellular

processes are analogous to the differential equations. In both WCMs growth is predicted by the continued feedback of the results of equations within biologically feasible bounds, which when met cause the cell to divide.

As WCMs can model genes in their genomic context, it is possible to use them to construct and test potential minimal gene sets, whereby genes are removed to form a proposed genome that cannot have any more genes removed from it without incurring loss of viability (Glass *et al.,* 2017). These gene sets have previously been proposed through protocell design (Tomita *et al.,* 1999), comparative genomics (Huang *et al.,* 2013), and single gene knockout studies (Karr *et al.,* 2012). Gene sets are quantified by combining candidate genes identified in these studies into comprehensive groups capable of accounting for all essential genes within a cell. Minimal genomes have also been proposed by running genetic algorithms GAMA and Minesweeper on the *M. genitalium* WCM (Rees-Garbutt *et al.,* 2020c). These algorithms select possible non-essential gene deletions before simulating the genome of *M. genitalium* without those genes and then analyse the resulting cell to check for viability. Simulations that produce dividing cells proceed to the next cycle, progressively increasing the number of genes deleted. This has allowed the automated production of 58,071 *M. genitalium* genomes (Rees-Garbutt *et al.,* 2020c) in a much more efficient manner than manual genome construction efforts. As individual gene analysis can impact the ability to meaningfully interpret results, researchers often use gene ontology (GO) terms, standardised labels that describe a gene's function, to reconnect the genes disrupted back to the processes being impacted (Apweiler *et al.,* 2004). GO terms are built from a structured, controlled vocabulary to create a network where each GO term is a node connected by the relationship between the terms, and as such can describe the relationships between genes and the processes they influence (Apweiler *et al.,* 2004). Rees-Garbutt *et al.* (2020c) utilise GO terms to analyse the minimal genomes produced by the GAMA and Minesweeper algorithms to identify the types of processes being dispensed with in the *in silico* cell.

1.5 - *E. coli* and *M. genitalium* as model and minimal organisms

Minimal genomes are sets of protein-coding genes that form a genome where no more genes can be removed without incurring loss of viability (Glass *et al.,* 2017). Such genomes can only exist in rich growth media with no external stressors (Hutchison *et al.,* 2016). The genes that cannot be dispensed with are therefore described as essential (Rancati *et al.,* 2018). In recent years essentiality has undergone a complex redefinition, resulting in the understanding that multiple minimal genomes for individual bacterial species can exist, depending on factors such as environmental conditions and which redundant genetic pathways are selected in the cell (Landon *et al.,* 2019; Xavier *et al.,* 2014; Rees-Garbutt *et al.,* 2020c). For example, low essential genes are dispensable in

most environmental and genetic contexts (Rancati *et al.,* 2018). There is disparity between finding a local minimum genome for a specific species and searching for a global minimum genome, which would contain the smallest number of genes necessary for life. A notable example of the former is a 38.9% gene reduction of *E. coli* (Iwadate *et al.,* 2011) while the latter is best represented by the creation of JCVI-Syn3.0*,* a 50% gene reduction of *Mycoplasma mycoides* (*M. mycoides*)*,* which possesses the smallest known genome to have ever autonomously produced a dividing cell (Hutchison *et al.,* 2016). Herein lies the divergence in strategy when producing a minimal genome. A number of previous genome reduction efforts are presented in Table 1.1. Many minimal genome creation efforts involve top-down construction, where extant genomes are reduced (Iwadate *et al.,* 2011). JCVI-Syn3.0 has a synthetic genome built in *Saccharomyces cerevisiae* and cloned into a *Mycoplasma* cell, which progresses towards a bottom-up approach (Göpfrich *et al.,* 2018). While routine bottom-up construction of an entire cell is currently infeasible due to complexity and costs, many synthetic biologists view this approach as an important way to construct the most minimal genomes (Forster & Church, 2006).

Table 1.1. An incomplete history of notable genome reduction efforts reported in the literature.

Year	Organism & Strain	Genome reduction	Reference
2002	*E. coli* MDS12	8%	Kolisnychenko *et al.*
2006	*E. coli* MDS43	15%	Pósfai *et al.*
2011	*E. coli* Δ33a	39%	Iwadate *et al.*
2013	*E. coli* DGF-298	35%	Hirokawa *et al.*
	Corynebacterium glutamicum MB001	6%	Baumgart *et al.*
2014	*E. coli* MS56	23%	Park *et al.*
	Pseudomonas putida EM383	4%	Martinez-Garcia *et al.*
2016	*M. mycoides* JCVI-syn3.0	50%	Hutchison *et al.*
	E. coli MGE-syn1.0 in *S. cerevisiae*	77%	Zhou *et al.*
	E. coli MDS69	20%	Karcagi *et al.*
	rE.coli-57	57-codon genome	Ostrov *et al.*
2017	*B. subitilis* PG10 and PS38	36%	Reuß *et al.*
2019	*E. coli* Syn61	synthetic 61-codon genome	Fredens *et al.*

It is important to note that reducing a genome to a state with the fewest genes is not the sole way of creating a minimal genome. Minimising the total genome size, including non-coding elements such as pseudogenes (Roberts & Morris, 2013) may provide avenues to further reduce a genome when no more genes are dispensable. Codon replacement research (Fredens *et al.,* 2019) has reduced the genome of *E. coli*, using an alternative metric to genes, by reducing the number of codons used to

synthesise proteins from DNA, providing us with another standard to measure minimality against. Mollicutes, of which *Mycoplasma* are a genus, have also been identified as candidates for the removal of non-essential enzymes (de Crécy-Lagard *et al.,* 2007), which would further decrease the contents of the cell. Progress in protein engineering has led to improvements in enzyme activity and stability, including colocalization efforts which have allowed us to understand where proteins occur and how they function (Li *et al.,* 2020). Theoretically, the ability to modify proteins and enzymes in this way could be scaled up to the entire proteome, streamlining cell metabolism and reducing complexity. A desirable direction for research could therefore be to combine efforts to reduce the genetic content of the cell with those to reduce the total number of components in the cell, producing the most minimal, least complex cell possible (Xavier *et al.,* 2014).

E. coli and *M. genitalium* serve as model and minimal organisms respectively (Taj *et al.,* 2014; Fraser *et al.,* 1995), and both organisms have been heavily used in minimisation efforts. The ease with which *E. coli* can be grown in the laboratory is a pertinent factor in its historical and continued use as a model organism for biological research, with quick doubling time, environmental resilience, and low cost of maintaining colonies all proving valuable characteristics (Blount, 2015). *M. genitalium* occupies an alternative niche in research, as it is notoriously difficult to culture effectively in the laboratory, with extremely slow reproduction frustrating research efforts (Hutchison *et al.,* 2016). This has stifled *in vivo* progress with *M. genitalium* and researchers have instead investigated other Mycoplasmas, for example the J. Craig Venter Institute (JCVI) synthesising the *M. mycoides* genome (Gibson *et al.,* 2010). Most recent research using *M. genitalium* has therefore been *in silico* (Rees-Garbutt *et al.,* 2020c).

The differing characteristics of *E. coli* and *M. genitalium* have prompted distinct avenues for genome design and minimisation. As an organism of specific interest to industry, utilised for large-scale biosynthesis of molecules such as terpenoids (Choi *et al.,* 2010), many previous attempts at minimising the genome of *E. coli* have focused on the removal of genes and metabolic pathways which do not overly diminish the rate of cellular growth and division on economically viable growth media (Hirokawa *et al.*, 2013). An example of such a minimised strain of *E. coli* is MDS42, representing a 14.3% genomic reduction compared to *E. coli* MG1655 without reducing growth rate (Pósfai *et al.,* 2006). This type of minimisation, while undeniably useful and representing progress in genome minimisation, is applied to an industrial setting where the goals are centred around the creation of chassis cells and does not explore the limits of genome minimisation investigated in other studies (Landon *et al*., 2019). The genomes of Mycoplasmas have undergone minimisation in its purest form; research efforts have created a strain of *M. genitalium* with the fewest genes possible for cell growth on ideal media, despite extended doubling time (Hutchison *et al.,* 2016). The

crowning achievement of this branch of research comes in the form of JCVI-syn3.0, a reduction of a previously synthesised *M. mycoides* genome which had been transplanted into *M. genitalium* and reduced to 473 genes (Hutchison *et al*., 2016). This organism represents the most genome-reduced cell produced to date, with a doubling time of ~180 minutes and providing exciting avenues for research regarding minimal gene sets, as well as further progress concerning universal core functions for life. JCVI-syn3A was subsequently developed by the JCVI, comprising 493 genes and producing a more robust cell which was easier to grow in the laboratory, with a doubling time of ~120 minutes (Hutchison *et al.,* 2016; Glass, 2017; Breuer *et al.,* 2019). Recent progress has been reported in JCVI-syn3A, as a reverse genetics approach revealed the seven genes required for normal cell division in this organism (Pelletier *et al.,* 2021).

1.6 - CRISPR-Cas9 and no-SCAR for bacterial genome engineering

Genome engineering has undergone huge revolutions in recent years, with the development of powerful technologies facilitating ever-growing control over our ability to manipulate the genomes of organisms (Khalil, 2020). Preceded by technologies such as zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) (Kim *et al.,* 1996; Boch *et al*., 2009), CRISPR-Cas9 technology represents arguably our most vital step forward in genome engineering (Jinek *et al.,* 2012). By utilising elements of the bacterial adaptive immune system, it is possible to target and edit genes with precision once thought impossible, enabling the vast editing of entire genomes to the accuracy of a few base pairs (Hsu *et al.,* 2014). This precision has significantly contributed to genetic studies conducted in eukaryotes, with achievements such as increasing rice grain yield through targeted gene knockouts (Lu *et al*., 2018).

As prokaryotes such as *E. coli* lack the ability to repair double-strand breaks (DSBs) in DNA by homologous recombination (HR), λ Red technology remains the premier method to conduct genome editing in *E. coli* (Murphy, 1998). Pioneered in 1998, λ Red recombineering for bacteria uses the bacteriophage λ Red proteins, Exo, Bet, and Gam to construct gene knockout mutants through HR of a double-stranded DNA (dsDNA) polymerase chain reaction (PCR) product with bacterial chromosomes (Murphy, 1998). Exo degrades dsDNA from 5’ ends to reveal the single-stranded DNA (ssDNA) which is targeted by Bet and HR takes place to insert the desired DNA, while Gam prevents the digestion of λ phage DNA so recombination can occur (Datsenko & Wanner, 2000). This revolutionary technique permitted the construction of the Keio collection, a single gene knockout library of *E. coli* K-12, amongst other achievements (Baba *et al.,* 2006; Mosberg *et al.,* 2010).

While previously frustrating the attempts of microbiologists to edit the genome of *E. coli*, the ability of CRISPR-Cas9 to induce cell death by creating DSBs now provides a way to optimise the process of

recombineering through counterselection. If a colony has not undergone successful transformation

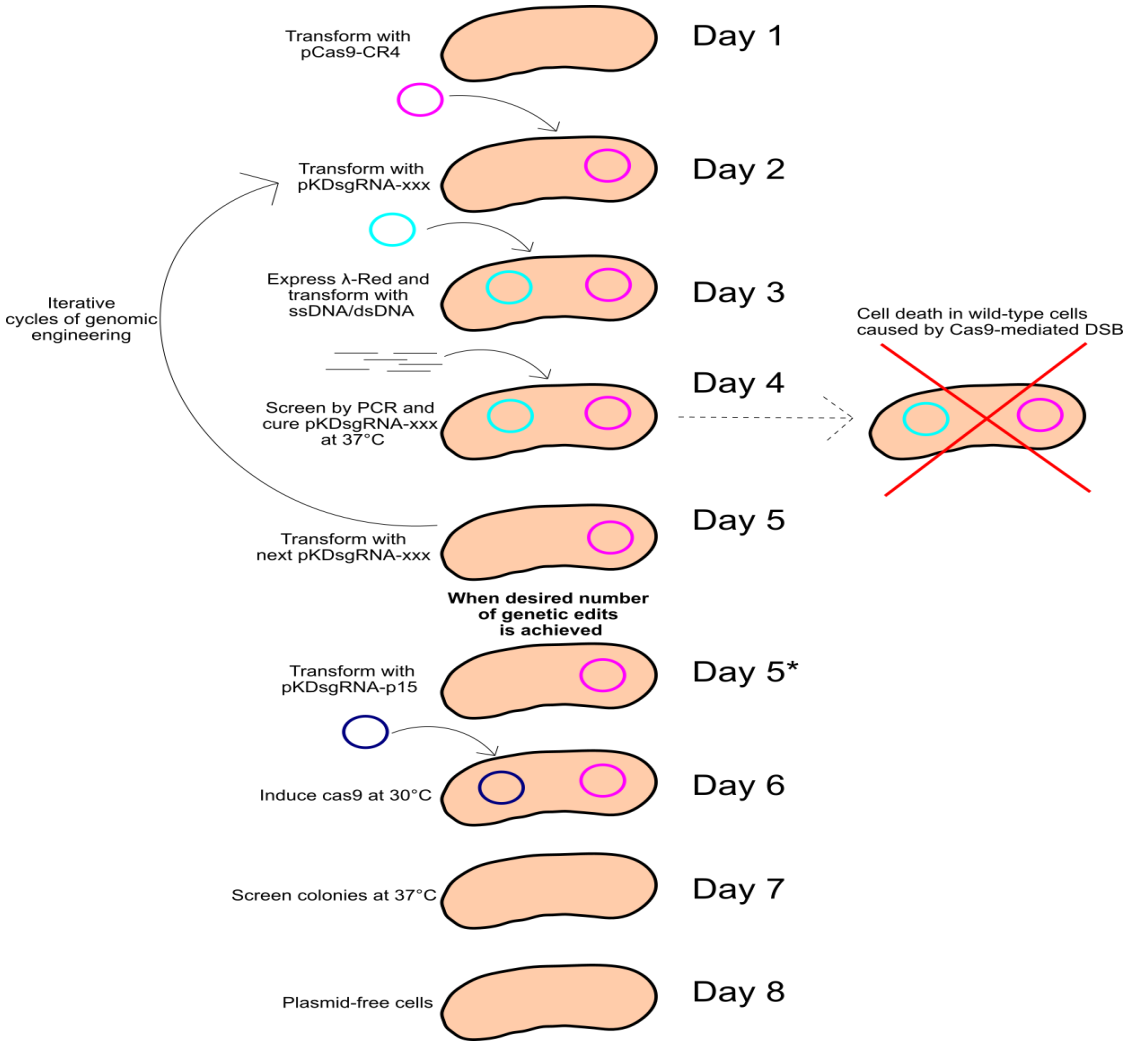

Figure 1.2. An outline of the no-SCAR method. Day 1: the pCas9-CR4 plasmid is transformed into electrocompetent *E. coli* and plated on LB + chloramphenicol at 37°C. Day 2: the pKDsgRNA-xxx plasmid (-xxx denotes the targeted gene) is transformed into cells and plated on LB + spectinomycin and chloramphenicol at 30°C. Day 3: cells are grown to OD ~0.5 in SOB then 50mM L-arabinose is added to induce λ Red. Cells are then made electrocompetent and transformed with ss/dsDNA to confer the desired mutation. Once recovered, cells are plated on LB + spectinomycin, chloramphenicol, and anhydrotetracycline at 30°C. Day 4: colonies are screened by PCR and grown at 37°C to cure the pKDsgRNA-xxx plasmid. Day 5: transform mutant cells with the next pKDsgRNA-xxx plasmid and repeat the process. Day 5*-8 curing: transform electrocompetent mutant cells with curing plasmid pKDsgRNA-p15 and recover cells in SOC at 30°C, add 100µg/L anhydrotetracycline at 30°C. Plate cells on LB + spectinomycin and anhydrotetracycline at 30°C. Screen colonies by patching on LB and LB + chloramphenicol at 37°C (adapted from Reisch & Prather, 2015).

and editing, then the wild-type cells will be successfully targeted by CRISPR and Cas9 will produce an irreparable DSB resulting in cell death (Bowater & Doherty, 2006). This phenomenon was capitalised upon to improve recombination efficiency in bacteria first by Jiang *et al.* (2013), and then by Reisch & Prather (2015). The latter reported the Scarless Cas9 Assisted Recombineering (no-SCAR) method of creating genomic edits in *E. coli* using λ Red recombineering, partnered with CRISPR-Cas9, to create scarless edits in a single step with relative ease (Reisch & Prather, 2015) (Figure 1.2). No-SCAR recombination works by employing two plasmids, one encoding single guide RNA (sgRNA) and λ Red machinery for recombineering by HR (pKDsgRNA-ack) and the other encoding a Cas9 nuclease to facilitate the creation of DSBs, thus inducing death in cells in which editing has not occurred (pCas9-CR4). A third, curing plasmid, is used to remove pCas9-CR4 from bacterial cells once all rounds of recombineering are complete, through introducing DSBs by Cas9 cleavage (pKDsgRNA-p15).

To target the desired genomic loci, pKDsgRNA-ack must be retargeted by Circular Polymerase Extension Cloning (CPEC) or round the horn cloning (Quan & Tian, 2011; Ochman *et al.,* 1988). To achieve cloning by CPEC, PCR is used to produce linear DNA fragments which have short overlapping sequences on both ends. The regions of overlap can be modified to target the desired genomic loci by changing the primers used for PCR. Once amplified, the products undergo *DpnI* digestion and are run on an agarose gel, before undergoing excision, purification, and further PCR cycling. The retargeted pKDsgRNA-ack is then transformed into electrocompetent *E. coli* by electroporation. The desired host strain can then be sequentially transformed with pCas9-CR4 and pKDsgRNA-ack, before transforming with linear DNA to create the desired deletion, insertion, or mutation in the bacterial genome. Here the λ Red machinery can be induced by L-arabinose to facilitate HR. Recovery and incubation after the editing takes place, alongside colony PCR and Sanger sequencing, will reveal whether editing was successful, as tetracycline-induced CRISPR-Cas9 mediated counterselection occurs. This system of recombineering can produce cyclical genome modifications with every iteration of editing increasing the amount of time saved, as pCas9-CR4 remains in cells. As such, it is possible to carry out three mutations in a cell and cure the plasmids used in 14 days, a method which is four days faster than SceI counter-selection (Reisch & Prather, 2015; Kim *et al.,* 2014). Reisch & Prather (2015) also reported high transformation efficiency, and low escape rate of counterselection by cells avoiding death by Cas9-induced DSBs, increasing the viability of the no-SCAR system for genome engineering.

1.7 - Aims

I set out to test the ability of the *E. coli* WCM to correctly predict the effect of gene knockouts. This would test how useful the model is for genome design. I also wanted to contribute to the research our group was conducting into *M. genitalium* minimal gene sets. I took three approaches:

1. To use the predictions of the *E. coli* WCM to inform and direct gene deletions in the laboratory using the no-SCAR system for genome editing (Reisch & Prather, 2015).
2. To help to identify and test possible minimal gene sets using the *M. genitalium* WCM.
3. To collaborate with the Covert group at Stanford University, the producers of both the *M. genitalium* and *E. coli* WCMs, to improve the *E. coli* WCM, which is still under development. Through this, I hope to increase the model's fidelity to living *E. coli* cells, allowing for more accurate predictions to be made using the model.

Chapter 2 – Minimal gene sets in *M. genitalium*

2.1 - Introduction

M. genitalium has been a vital subject of minimal genome research (see Chapter 1 section 1.4). *M. genitalium's* reduced metabolism and simplicity made it possible to create a WCM of it, that has allowed *in silico* investigation of its genome (Karr *et al.,* 2012). Our research group has been working with the *M. genitalium* WCM for a number of years, with it being the target of the genome design algorithms GAMA and Minesweeper (Rees-Garbutt *et al.,* 2020c). These algorithms permit the use of design-simulate-test cycles for genome design and they were first applied to genome minimisation in *M. genitalium* (Rees-Garbutt *et al.,* 2020c). In addition to this work, members of our research group utilised their knowledge of the *M. genitalium* WCM to investigate the potential minimal gene set for life (Rees-Garbutt *et al.,* 2020b). Many researchers have published gene sets that they claim should function as bacterial minimal gene sets, but these studies were based on removing one or at most a few genes at a time and then combining genes based on the results, assuming that there would be no cumulative effects of removing large numbers of genes simultaneously (Forster & Church, 2006; Gil, 2014; Glass *et al.,* 2006; Huang *et al*., 2013; Hutchison *et al.,* 1999; Karr *et al.,* 2012; Mushegian & Koonin, 1996; Tomita *et al.,* 1999). Joshua Rees-Garbutt realised that we could test these gene sets in the *M. genitalium* WCM to see whether they could produce dividing cells *in silico*, as this would be practically impossible to investigate *in vivo*. Despite *in silico* research using the *M. genitalium* WCM being faster than *in vivo* research, there are significant time costs associated with the model's use. The simulations which produced the data I analyse in this chapter each took between 5 – 12 hours, increasing the proportion of time for my project which was dedicated to this work. Joshua Rees-Garbutt discovered that none of the minimal gene sets chosen for investigation produced dividing *in silico* cells, with each set requiring reintroduction of specific essential and low-essential genes to enable cellular division (Rees-Garbutt *et al.,* 2020b).

Table 2.1 details which minimal gene sets were investigated and what they were called for our research. The processes used to design the gene sets are also given (e.g. protocell creation, comparative genomics, and single gene deletions). Protocells are collections of biomolecules encapsulated in a membrane and are capable of self-replication (Dzieciol & Mann, 2012). Comparative genomics computationally and experimentally compares the genomes of different species to identify common genes (Koonin, 2003). Table 2.2 is a summary of results from Rees-Garbutt *et al.* (2020b), identifying the sizes of the initial and final *in silico* genomes for the minimal gene sets investigated. The classes of genes (see Chapter 1 section 1.4) which had to be reintroduced are also presented. Rees-Garbutt *et al.* (2020c) previously published *M. genitalium*

genomes *in silico* with sizes of 256 (Minesweeper_256) and 236 (GAMA_236) genes, which remain the most minimal genomes shown to divide *in silico* to date.

Table 2.1. Comparing minimal gene sets with *M. genitalium in vivo* and the *M. genitalium* whole-cell model. Predicted *in silico* genome size is the size of gene sets once adapted for use in the *M. genitalium* WCM, with unmodelled genes removed and 42 RNA coding genes added. * = *protein-coding genes. M. genitalium has 42 RNA-coding genes that are not included in this column.* ^ = *due to unknown function* (adapted from Rees-Garbutt *et al.,* 2020b).

Source of minimal gene set	Code name in Rees-Garbutt et al. (2020b)	Code name in this thesis	Design approach	*In vivo* genome design size*	Unmodelled genes^ in genome design	Unmodelled genes^ in gene deletions	Predicted *in silico* genome size
Forster and Church 2006	Nashville	Church	Protocell	89	0	-	131
Tomita *et al.* 1999	Fujisawa	Tomita	Protocell	98	0	-	140
Huang *et al.* 2013	Guelph	Huang	Comparative Genomics	123	5	-	160
Gil 2014	Valencia	Gil	Comparative Genomics	180	6	-	216
Mushegian and Koonin 1996	Bethesda	Koonin	Comparative Genomics	253	12	-	283
Karr *et al.* 2012	Stanford	Karr	Single Gene Deletions	242	-	0	284
Glass *et al.* 2006	Rockville 2	Glass	Single Gene Deletions	258	-	44	344
Hutchison *et al. 1999*	Rockville	Hutchison	Single Gene Deletions	265	-	41	348
-	*M. genitalium* whole-cell model	-	-	359	124	-	401

My contribution to this research was to analyse the minimal gene sets to try to identify biological functions that could be dispensed with while still producing a dividing cell. To do this I took the *in silico* simulations and grouped genes according to function using GO terms. I also performed GO term analysis of the genes Joshua Rees-Garbutt reintroduced to the gene sets sourced from the literature in order to produce cells which successfully divided *in silico*. This work is currently undergoing review after submission for potential publication in ACS Synthetic Biology.

Table 2.2. Minimal gene sets before and after the reintroduction of essential and low essential genes to produce dividing in *in silico* cells. The "Agreed" set contains 14 genes not included in all eight minimal gene sets. * = *Final size of in silico genome includes members of the original 359 protein-coding genes and the 42 RNA-coding genes in the M. genitalium whole-cell model, to realistically indicate the genome that might need to be built to produce living cells* (adapted from Rees-Garbutt *et al.,* 2020b).

Gene set code name in Rees-Garbutt et al. (2020b)	*In silico* gene deletions (cells did not divide)	*In silico* gene deletions (cells divided)	Essential genes reintroduced	Low essential genes reintroduced	Final size of *in silico* genome*
Nashville	270	142	121	7	259
Fujisawa	261	142	112	7	259
Guelph	241	128	107	6	273
Valencia	185	110	69	6	291
Stanford	117	109	3	5	292
Bethesda	118	82	34	2	319
Rockville 2	57	45	9	3	356
Rockville	53	43	9	1	358
Agreed	14	13	0	1	388

2.2 - Methods

2.2.1 - Data sources

The *M. genitalium* WCM is publicly available at: https://github.com/CovertLab/WholeCell. The scripts and data I used to analyse the minimal gene sets are available on GitHub (https://github.com/JJRightmyer/MGS-processing and https://github.com/squishybinary/Gene_Ontology_Comparison_for_Mycoplasma_genitalium_whole-cell_model).

2.2.2 - Procedures

2.2.2.1 - GO term analysis of minimal gene sets

Having received the gene sets from Joshua Rees-Garbutt, I created .txt files of the genes present in each gene set. This resulted in the creation of eight .txt files, one for each minimal gene set: ChurchMGS.txt (Forster & Church, 2006), GilMGS.txt (Gil, 2014), GlassMGS.txt (Glass *et al.,* 2006), HuangMGS.txt (Huang *et al*., 2013), HutchisonMGS.txt (Hutchison *et al.,* 1999), KarrMGS.txt (Karr *et al.,* 2012), KooninMGS.txt (Mushegian & Koonin, 1996), and TomitaMGS.txt (Tomita *et al.,* 1999). These files are collectively referred to as AuthorNameMGS.txt. The GO terms were retrieved from UniProt by Joshua Rees-Garbutt (Apweiler *et al.,* 2004) and I processed results using a script in Jupyter Notebooks created by Joshua Rees-Garbutt (https://github.com/squishybinary/Gene_Ontology_Comparison_for_Mycoplasma_genitalium_whole-cell_model/blob/master/165deletions_versus_GOBaseline.ipynb) which converted the list of genes (AuthorNameMGS.txt) knocked out from the model to create each minimal gene set into .csv

files for analysis (AuthorNameMGSvsGOBaseline.txt). To create the .csv files I made the following changes to the code in file 165deletions_versus_GOBaseline.ipynb:

1. Change line 4 to read in the gene set-specific .txt file (AuthorNameMGS.txt).

   ```
   df2 = pd.read_csv('165deletions.txt', sep="\t")
   ```

   To

   ```
   df2 = pd.read_csv('AuthorNameMGS.txt', sep="\t")
   ```

2. Change line 9 to output a .csv file specific to the gene set that was inputted.

   ```
   Df3.to_csv('165delsvsGOBaseline.txt', sep='\t'
   ```

   To

   ```
   df3.to_csv('AuthorNameMGSvsGOBaseline.txt', sep='\t'
   ```

This process involved reading in the GO baseline *M. genitalium* data which each minimal gene set would be compared to (GOBaseline.txt), then reading in the gene deletion list (AuthorNameMGS.txt) as a 1D vertical dataframe and converting it to a list. The deletion list for all eight minimal gene sets was read through repeatedly running the code (ChurchMGS.txt, GilMGS.txt, GlassMGS.txt, HuangMGS.txt, HutchisonMGS.txt, KarrMGS.txt, KooninMGS.txt, and TomitaMGS.txt). The baseline and deletion list would then be compared in a dataframe, with GeneIDs present in the baseline and not in the deletion list being kept in the new list. This list was then exported to a .csv file. I then manually organised GO terms by whether they were unaffected, reduced, or removed by gene deletions (Appendices 1-8). I proceeded to analyse the tables and synthesise information from across the gene sets to identify patterns and trends.

2.2.2.2 - GO term reintroductions for minimal gene sets

I repeated the process outlined in Section 2.2.2.1 for genes which when reintroduced produced minimal gene sets that divided *in silico*, with the only difference being GeneIDs present in both the baseline and the reintroduction list were kept in the new list, which was exported to a .csv file. Joshua Rees-Garbutt adapted the 165deletions_versus_GOBaseline.ipynb script to create Reintroductions_Protocells_versus_GOBaseline.ipynb (https://github.com/JJRightmyer/MGS-processing/blob/main/Reintroductions_Protocells_versus_GOBaseline.ipynb) with the following changes:

1. Change line 4 to read in the reintroduction gene list .txt file.

   ```
   Df2 = pd.read_csv('165deletions.txt', sep="\t")
   ```

   To

```
df2 = pd.read_csv('reintroduction_authorName.txt', sep="\t")
```

2. Change line 7 to print the reintroduction list as opposed to the deletion list.

```
Print(deletionlist)
```

To

```
print(reintrolist)
```

3. Change lines 8 and 9 so the dataframe constructed compares the baseline and reintroduction lists, then export that list to a .csv file.

```
Df3 = df[~df['GeneID'].isin(deletionlist)]
df3.to_csv('165delsvsGOBaseline.txt', sep='\t')
```

To

```
df3 = df[df['GeneID'].isin(reintrolist)]
df3.to_csv('reintroductionAuthorNameGOList.txt', sep='\t')
```

Once I had created the .csv files comparing the reintroduced GO terms to the baseline for *M. genitalium* by processing each of the eight gene reintroductions through Reintroductions_Protocells_versus_GOBaseline.ipynb, I manually produced tables to allow for easy comparison between the gene sets and looked for patterns across the data (Appendices 9-16).

2.3 - Results

2.3.1 - GO term reductions and removals

Through grouping *M. genitalium* genes using GO terms, it was possible to reveal more about the cellular functions that were either being conserved or dispensed with in the different minimal gene sets. When compared to the baseline genome (all genes present, 347 GO term categories), the minimal gene sets each provided a unique combination of GO term reductions and removals (Table 2.3). For brevity, the complete tables of GO terms for all eight gene sets are included in Appendices 1-8, rather than here.

Table 2.3. A summary of minimal gene set (MGS) GO term reductions and removals from the baseline 155 GO terms. The minimal gene set names correspond to the publications identified in Table 2.1.

MGS	GO terms reduced	GO terms removed
Church	22	40
Gil	21	30
Glass	17	14
Huang	24	36
Hutchison	15	8
Karr	18	29
Koonin	19	18
Tomita	22	39

The Church MGS (Appendix 1) represented one of the most reduced genomes, with reductions and removals in GO categories across the entire genome. Processes affected included: DNA (replication, topology, transcription, recombination, repair, catabolism, nucleotide synthesis), RNA (processing, modification, compound salvage, pseudouridine synthesis), ATP and important biomolecule metabolism and regulation (carbohydrates, sugars, lipids, carboxylic acids, ATP coupled proton transport, coenzyme A biosynthesis, fructose 1,6-bisphosphate metabolic process), environmental interactions (pathogenesis, cell adhesion, cytoadherence to microvasculature, protein secretion, heterophilic cell-cell adhesion), cell cycle (cell division, chromosome segregation & separation), protein-related processes (transport, lipoylation, ribosome biogenesis, lipoprotein, catabolism, import repair, modification), and homeostasis (response to oxidative stress, cellular phosphate ion homeostasis). These reductions and removals represented a system-wide reduction in environmental interaction, damage repair, energy production, and process regulation. The removal of various molecule transport mechanisms would also lead to increased generation time, as reactions would be slowed down. It is likely that a cell with this genome would be very susceptible to environmental change as well as deleterious mutations, as four of its DNA repair systems have been heavily reduced or removed.

The Gil MGS (Appendix 2) had significant reduction, but not to the same degree as MGSs Church, Tomita, and Huang. Processes conserved include RNA processes, processes involved in the production of ATP, and processes related to proteins such as transport and folding. Otherwise, the reductions were very similar to MGSs Church, Tomita, and Huang, with genes involved in DNA processes, environmental interaction, cell cycle, and homeostatic processes identical to the Huang MGS. This set would possibly have increased protein availability and functionality over other gene

sets due to tRNA processing and protein folding being unchanged, as well as increased ATP production in the cell. While the Gil MGS introduces no unique reductions or removals of GO terms, its composition aids in identifying a general base for the biological functions which can be dispensed with whilst still producing a dividing *in silico* cell. Decreases in the cell's ability to interact with its environment were generally tolerated by gene sets produced by protocell and comparative genomic studies (MGSs Church, Tomita, Huang, Gil, and Koonin) but not by gene sets produced by single gene deletions (MGSs Karr, Glass, and Hutchison).

The Glass MGS (Appendix 3) had significant conservation of processes across the genome, comparable to that of the Hutchison set. While all GO categories were still reduced, only the cell cycle was reduced a significant amount, with reductions equal to MGSs Koonin, Karr, and Gil. The gene set also showed reduction to four DNA repair processes, possibly reducing the cell's ability to tolerate genetic damage. However, it retained full interaction with the environment, with the DNA restriction-modification system being the only process removed. This is also the only gene set apart from MGS Hutchison to conserve cellular phosphate ion homeostasis, with all others either reducing or removing the process.

Another heavily reduced genome, the Huang MGS (Appendix 4) was very similar to MGSs Church and Tomita, including protein folding while excluding fructose 1,6-bisphosphate metabolism and some other protein-related processes. The Huang MGS also removed sister chromatid cohesion and chromosome condensation, both with consequences for cell division. This could increase generation time as it becomes harder to package chromosome copies into daughter cells.

The Hutchison MGS (Appendix 5) showed the capacity for the reduction of GO categories including DNA (repair, transcription, recombination), RNA (tRNA processing, mRNA catabolic, pseudouridine synthesis), ribosome biogenesis (with consequences for gene translation), environmental interaction (pathogenesis, cell adhesion, cytoadherence to microvasculature), cell cycle (division, chromosome segregation), and ATP generation including metabolism and regulation of important biomolecules (carbohydrates, sugars, lipids, coenzyme A biosynthetic process, phosphate ion transmembrane transport). In many cases analogous pathways exist to compensate for the removal of certain processes, as well as the ideal growth environment negating the need for some environmental interaction processes. While the gene set would remove redundancies, it is also important to note the removal of multiple DNA repair mechanisms may impact the survival of cells across generations as deleterious mutations accumulate.

The Karr MGS (Appendix 6) had GO term process reduction which fell between the heavily reduced sets and the lightly reduced. While maintaining reductions in ATP and important biomolecule

metabolism comparable to MGSs Huang and Tomita, as well as cell cycle reductions identical to MGSs Huang and Gil, the Karr MGS saw significant process conservation across RNA processes, environmental interactions, and protein-related processes. This represented a prominent set of environmental interaction processes, which have often been either reduced or removed in other gene sets. This would likely create a cell that is more adaptable to environmental change, as well as being able to produce more functional proteins, as relatively few protein-related processes are affected.

The Koonin MGS (Appendix 7) was relatively lightly reduced, with significant process conservation compared to the heavily reduced sets in DNA, RNA, ATP and important biomolecule metabolism, and protein-related processes. Its interactions with the environment, cell cycle, and homeostatic processes were as reduced as the most minimal gene sets. Uniquely, this gene set only reduced one DNA repair process, making the cell relatively robust to genetic damage.

The Tomita MGS (Appendix 8) was very similar to the Church MGS, with only the inclusion of protein folding and refolding and the exclusion of glycolytic process, carboxylic acid metabolism, and fructose 1,6-bisphosphate metabolic process setting them apart. This would result in the cell behaving in much the same way as the Church MGS, with low tolerance to environmental or genetic stress. By impacting protein folding and refolding, it is likely that this cell will have reduced rates of cellular reactions, due to enzymes not having the correct active site.

2.3.2 - Gene reintroductions

GO term analysis of the genes reintroduced to *M. genitalium* gene sets was also successful, clarifying which cellular functions the *in silico* cells could not dispense with when simulated using the *M. genitalium* WCM. The variation in required GO term reintroductions to produce dividing *in silico* cells was immediately apparent, with the Karr MGS requiring the fewest reintroductions by far (Table 2.4). This result is unsurprising as the Karr MGS was proposed by the same authors who developed the *M. genitalium* WCM. Complete tables of all the GO terms reintroduced to produce functioning *in silico* cells are included in Appendices 9-16.

Table 2.4. A summary of minimal gene set (MGS) GO term reintroductions. Both the number of categories reintroduced and the number of processes that make up those categories are included.

MGS	GO term categories reintroduced	GO term processes reintroduced
Church	91	153
Gil	50	73
Glass	14	15
Huang	65	115
Hutchison	12	13
Karr	1	2
Koonin	25	35
Tomita	80	130

Of the GO terms which had to be reintroduced to produce dividing *in silico* cells, translation, transport, tRNA processing, and glycolytic process were most common, with at least one appearing in each of the gene sets. DNA replication, transcription DNA-templated, and carbohydrate metabolic process were also reintroduced often. The Karr MGS was notable for only requiring the reintroduction of genes which were involved in transport to produce a dividing *in silico* cell. The trends of functions restored to the gene sets once genes were reintroduced are discussed in more detail by Rees-Garbutt *et al.* (2020b), where common cellular functions which required reintroduction are identified. Reintroduced cellular functions identified by Rees-Garbutt *et al.* (2020b) included: DNA (polymerase subunits, thymidine insertion, recycling of pyrimidine, chromosome segregation; RNA (polymerase subunit, tRNA modification, the 50S and 30S ribosomal subunits); transporters (cobalt, phosphonate, potassium); production (NAD, flavin, NADP, fatty acid/phospholipids); and dehydrogenation (glycerol and alpha-keto acids). The reintroduction of phosphonate transport (MG_291) was vital to produce dividing *in silico* cells, as this process was removed in all MGSs bar Koonin, depriving *in silico* cells of phosphate. Phosphonate transport had to be reintroduced to ensure *in silico* cells divided for the other seven MGSs.

2.4 - Discussion

These results clarify which biological functions should be included in any minimal gene set for *M. genitalium*. They identify which mistakes previous authors had made when proposing their own gene sets and provide useful information for future genome designers. GO term analysis permits contextualisation of the genes that are both conserved and dispensed with to produce a viable minimal genome for *M. genitalium*, assigning cellular processes to each gene in question, allowing us to understand how the *in silico* cell is being impacted when its genome is reduced. As published in Rees-Garbutt *et al.* (2020b), MGSs Church and Tomita are the smallest *in silico* genomes shown to function behind Minesweeper_256 and GAMA_236 (Rees-Garbutt *et al.,* 2020c), which is largely

reflected by the GO term analysis of these gene sets (Table 2.3). These gene sets needed the most gene reintroductions to produce dividing *in silico* cells, and reintroduction of the most GO terms (Table 2.4). Through the identification of cellular processes deemed essential to produce viable *M. genitalium* cells, for example the necessity of having processes in place to provide cells with sufficient phosphate (section 2.3.2), we can provide a more comprehensive blueprint to direct future *in vivo* construction of minimal genomes. While WCM *M. genitalium* cannot perfectly replicate a cell's response to genetic manipulation (see Chapter 1 section 1.3), this research nonetheless progresses minimal genome design by improving upon potential MGSs presented in the literature and providing a clear direction for future research. With the development of more efficient technologies permitting reconstruction of *Mycoplasma* genomes in yeast cells (Benders *et al.,* 2010; Karas *et al.,* 2013; Tsarmpopoulos *et al.,* 2016), synthetic biologists have the necessary tools to begin investigating the MGSs we present *in vivo*.

Chapter 3 – Laboratory work informed by WCM predictions

3.1 - Introduction

To comprehensively assess the accuracy of WCMs, we must test their predictions *in vivo*. Only through *in vivo* investigation is it possible to determine how genome design methodologies can be improved by using WCMs. The development of the *E. coli* WCM provides us with the chance to investigate WCM predictions *in vivo* in an efficient and cost-effective manner, as *E. coli* is much easier to work with in the laboratory than *M. genitalium* (see Chapter 1 section 1.4). By implementing a system for genome editing of *E. coli* in our laboratories, we would be able to test the *in silico* predictions of the *E. coli* WCM to better understand how accurately it relates to *in vivo* cells.

While our group had access to the *E. coli* WCM, it was designed to run on Stanford University's SHERLOCK supercomputer cluster, so was initially incompatible with the University of Bristol's BlueCrystal supercomputer cluster. To get it running successfully required assistance from the Advanced Computing Research Centre team at the University of Bristol. Whilst other team members worked on this and to make efficient use of time, I decided to set up the no-SCAR protocol (see Chapter 1 section 1.5) in our laboratory and trial a proof-of-concept deletion so that the protocol would be ready to use for *in vivo* investigations of *in silico* results once we could produce them (Reisch & Prather, 2015). When fully implemented, the no-SCAR protocol allows for an initial genomic edit to be created in less than a week, with consequent iterations possible to complete in three or four days, assuming an efficient workflow with plasmid retargeting and oligonucleotide creation. This is one of the fastest methods for iterative, low error rate genome engineering, which is ideal as the laboratory work would likely be the limiting step on data collection, as *in silico* results with the *E.* coli WCM can be produced in as little as 15 minutes.

3.2 - Methods

Here I document the laboratory methods I undertook as I attempted to implement the no-SCAR method of genome engineering (Figure 1.2). I also detail the steps I would have taken if the effects of the Covid-19 pandemic had not caused our laboratories to close.

3.2.1 - Strains and constructs

E. coli K-12 substrain MG1655 (Genotype: *F- λ- ilvG- rfb-50 rph-1*; Blattner *et al.,* 1997) was obtained as frozen stocks from Thomas Gorochowski, University of Bristol. Frozen bacterial stocks were stored in an equal volume of 50% glycerol solution (VWR, 444485B) gently mixed with overnight culture and stored at -80°C.

I received plasmids pCas9-CR4, pKDsgRNA-ack, and pKDsgRNA-p15 from Addgene in the form of *E. coli* DH5-α (Genotype: F^- *endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG purB20* ϕ80d*lacZ*ΔM15 Δ(*lacZYA-argF*)U169, hsdR17($r_K^- m_K^+$), λ^-; Hanahan, 1985) bacterial stabs, from which I made glycerol stocks and stored at -80°C. The no-SCAR plasmids pKDsgRNA-ack (Addgene plasmid # 62654 ; http://n2t.net/32ristol:62654 ; RRID:Addgene_62654), pCas9-CR4 (Addgene plasmid # 62655 ; http://n2t.net/32ristol:62655 ; RRID:Addgene_62655), and pKDsgRNA-p15 (Addgene plasmid # 62656 ; http://n2t.net/32ristol:62656 ; RRID:Addgene_62656) were gifts from Kristala Prather (Reisch & Prather, 2015). Plasmid maps for plasmids pCas9-CR4 (https://www.addgene.org/62655/), pKDsgRNA-ack (https://www.addgene.org/62654/), and pKDsgRNA-p15 (https://www.addgene.org/62656/) are available at Addgene. Cultures containing pCas9-CR4 were grown on spectinomycin at 37°C while those containing pKDsgRNA-ack and pKDsgRNA-p15 were grown on chloramphenicol at 30°C.

3.2.2 - Bacterial cultures

Fresh bacterial cultures were prepared by streaking from frozen stocks to single colonies on Luria-Bertani (Miller) broth (LB) (Supelco, L3027-250G) 1% agar plates (Sarstedt, 82.1472). Plates were sealed with parafilm and incubated at 37°C overnight. Duplicate liquid cultures were prepared by inoculating 5ml LB media (Applichem, A6666.0500G) in a 15ml Falcon tube, with a single picked colony from the LB agar plates. A duplicate tube of medium was set up but not inoculated as a control to check the quality of my aseptic technique. The Falcon tubes were incubated overnight at 37°C and 250rpm.

3.2.3 - Preparing electrocompetent cells

Overnight culture was diluted into 5ml of fresh LB media in a 15ml Falcon culture tube to an OD_{600} of 0.1, then incubated it at 37°C and 250rpm until an OD_{600} of 0.6 was reached. Cells were harvested by centrifugation at 4°C, 4100 x g for 20 minutes, the supernatant discarded and the pellet resuspended in 5ml ice-cold, sterile MilliQ water before centrifuging again at 4°C, 4100 x g for 10 minutes. This was repeated three times to wash the cells, followed by a fourth wash of the pellet with 5ml ice-cold, sterile 10% glycerol, centrifuged at 4°C, 4100 x g for 10 minutes. The supernatant was discarded and the electrocompetent cells were resuspended in 50µl ice-cold, sterile 10% glycerol, divided into two 25µl aliquots and frozen at -80°C.

3.2.4 - Confirming plasmid structure

Plasmid structures were checked by culturing them in host bacteria under the appropriate antibiotic selection, isolating plasmid DNA and digesting using restriction enzymes.

3.2.4.1 - Antibiotics

Antibiotic stock solutions were made by diluting 50mg chloramphenicol powder (Sigma-Aldrich, C0378-25G) in 2ml ethanol (VWR, MFCD00003568) to produce a stock solution with a concentration of 25mg/ml, and diluting 100mg spectinomycin powder (Cambridge Bioscience, S016-5g) in 2ml dH_2O to produce a stock solution with a concentration of 50mg/ml.

3.2.4.2 - Selective growth

Chloramphenicol and spectinomycin stock solutions (section 3.2.4.1) were diluted to 25µg/ml and 50µg/ml respectively in 1% LB agar solution to make six plates (section 3.2.2). Three plates contained chloramphenicol and the other three spectinomycin. Cultures containing plasmids pCas9-CR4, pKDsgRNA-ack, and pKDsgRNA-p15 were grown on the antibiotic they were resistant to (section 3.2.1) alongside duplicate cultures as negative controls.

3.2.4.3 - Restriction digests

Restriction digests were conducted using restriction enzymes EcoRI (Promega, R6011) and XhoI (Promega, R6161) following the Promega Assembly of Restriction Enzyme Digestions manual (https://www.promega.com/-/media/files/resources/protocols/technical-manuals/101/restriction-enzymes-protocol.pdf). Plasmids were first miniprepped using the QIAprep Spin Miniprep Kit (50) (QIAGEN, 27104), then digested before undergoing electrophoresis on 1% agarose gel. Uncut plasmids were run alongside digested plasmids as a control to check the success of restriction digests.

3.2.4.4 - Agarose gel electrophoresis

1% agarose gels (Fisher Scientific, 13478388) were prepared for electrophoresis using MidoriGreen dye (Nippon Genetics, MG05), NEB Quick-Load® Purple 1 kb Plus DNA Ladder (NEB, N0550S) and Quick-Load® 100 bp DNA Ladder (NEB, N0467S). The gels ran for 120 minutes at 80V.

3.3 - Results

3.3.1 - Plasmid validation by selective growth

To validate the plasmids pCas9-CR4, pKDsgRNA-ack, and pKDsgRNA-p15, I grew cultures containing plasmids on selective media (section 3.2.4.2). Three plates contained chloramphenicol as a selective agent, of which two were used to grow cells containing pKDsgRNA-ack and pKDsgRNA-p15 plasmids, which confer chloramphenicol resistance. The other served as a control for growing cells containing the pCas9-CR4 plasmid, which confers spectinomycin resistance. The other three plates contained spectinomycin, one to select for cells containing the pCas9-CR4 plasmid, and the others to act as controls for the pKDsgRNA-ack and pKDsgRNA-p15 plasmids. The cultures all grew on the media they

had respective resistances to, and the negative controls were all successful, indicating presence of the correct plasmids.

3.3.2 - Plasmid validation by restriction digestion

I wanted to further clarify the structures of the no-SCAR plasmids, as to be certain of their effects when used in recombineering. I achieved this by restriction digest (section 3.2.4.3). I used Addgene's Sequence Analyser tool to find the cut sites of enzymes on all three plasmids when using restriction enzymes EcoRI and XhoI. For plasmids pKDsgRNA-ack and pKDsgRNA-p15, EcoRI cuts at 1215bp, 2724bp, and 5275bp, while XhoI cuts once at 5816bp (https://www.addgene.org/browse/sequence/258350/; https://www.addgene.org/browse/sequence/228332/). For plasmid pCas9-CR4, EcoRI cuts at 739bp and 2111bp, while XhoI cuts once at 4912bp (https://www.addgene.org/browse/sequence/212888/). Once extracted and clean, I separately incubated the plasmids with both EcoRI and XhoI (section 3.2.4.3). EcoRI should produce bands of 1509bp, 2551bp, and 2868bp for plasmids pKDsgRNA-ack and pKDsgRNA-p15, and bands of 1372bp and 5398bp for plasmid pCas9-CR4. XhoI should produce bands of 6959bp and 6928bp for plasmids pKDsgRNA-ack and pKDsgRNA-p15 respectively and produce a 6770bp band for plasmid pCas9-CR4. I made and ran two 1% agarose gels with wells for electrophoresis (section 3.2.4.4). This digest proved problematic, with no clear bands being produced by any restriction enzymes, while the uncut plasmid produced clear bands of the expected size.

3.3.3 - Restriction digest troubleshooting

I proceeded to troubleshoot this digest through individually altering reagents and conditions to discover what was causing the digest to behave unexpectedly. The dH_2O I was using in the restriction digestion reaction (Table 3.1) could have been contaminated with DNases which were completely digesting the plasmids. I autoclaved Milli-Q water before attempting another restriction digest, but again the digests produced no clear bands when run on 1% agarose gel. As suggested by Colin Lazarus, University of Bristol, I also tried incubating my miniprepped DNA at 70°C for 10 minutes before the addition of restriction enzymes to denature exonucleases which may have been present and interfering with the digest, but this did not help. Further research suggested that using enzymes at a concentration which was too high would result in errant digestion of plasmid DNA, producing fragments of unexpected sizes (https://www.thermofisher.com/uk/en/home/life-science/cloning/cloning-learning-center/34ristol34en-school-of-molecular-biology/molecular-cloning/restriction-enzymes/restriction-enzyme-key-considerations.html). I tried reducing the volume of restriction enzyme in solution to 0.2µl. As this did not produce clear banding on 1% agarose gel, I decided to closely analyse the concentrations of the entire restriction enzyme solution.

Using a nanodrop to discern the exact concentration of plasmid DNA produced by miniprepping the plasmids, I took the minipreps with the highest concentrations (pCas9-CR4: 91.40ng/µl, pKDsgRNA-ack: 51.70ng/µl, pKDsgRNA-p15: 48.50ng/µl) of DNA to be digested. This was to ensure there was enough DNA in the digestion so that it would both be visible on a 1% agarose gel and to avoid nonspecific degradation of DNA by star activity. To ensure I achieved the correct ratio of DNA to restriction enzyme to maximise the chance of a successful restriction digest, I adjusted the volumes of the DNA solution and restriction enzyme solution in the reaction, also changing the volume of dH_2O I used. The changes I made are summarised in Table 3.1.

Table 3.1. The composition of restriction digestion reactions, comparing my initial unsuccessful effort to the subsequently adjusted, and successful effort below. DNA concentrations for the successful run were: pCas9-CR4: 91.40ng/µl, pKDsgRNA-ack: 51.70ng/µl, pKDsgRNA-p15: 48.50ng/µl.

Plasmid	Restriction Enyzme 10X Buffer (µl)	Bovine Serum Albumin (µl)	DNA solution (µl)	Restriction Enzyme (µl)	dH_2O (µl)	Result
pCas9-CR4	2.00	0.20	1.00	0.50	16.30	Unsuccessful
pKDsgRNA-ack	2.00	0.20	1.00	0.50	16.30	Unsuccessful
pKDsgRNA-p15	2.00	0.20	1.00	0.50	16.30	Unsuccessful
pCas9-CR4	2.00	0.20	3.00	0.12	14.68	Successful
pKDsgRNA-ack	2.00	0.20	4.84	0.12	12.84	Successful
pKDsgRNA-p15	2.00	0.20	5.16	0.12	12.52	Successful

This restriction digest was largely successful, with agarose gel electrophoresis producing bands of the expected size for many combinations of restriction enzyme. Digestion of plasmid pCas9-CR4 with EcoRI produced a band consistent with the prediction of 5398bp but produced a band larger than the expected 1372bp. I would have investigated this further if I had more time in the laboratory. Digestion with XhoI produced a band consistent with the 6770bp prediction. Digestion of plasmid pKDsgRNA-p15 with EcoRI produced bands consistent with predicted bands of 1509bp, 2551bp, and 2868bp, while digestion with XhoI produced a band consistent with the 6928bp prediction (Figure 3.1). Unfortunately, the capture of the agarose gel showing the restriction digest of plasmid pKDsgRNA-ack did not successfully download and due to Covid-19 closing our laboratories, I could not return to repeat the gel electrophoresis. I was able to note down the results of this gel

electrophoresis and saw that digestion with EcoRI produced bands consistent with the predicted sizes of 1509bp, 2551bp, and 2868bp for plasmid pKDsgRNA-ack, as well as a 6959bp band when digested with XhoI.

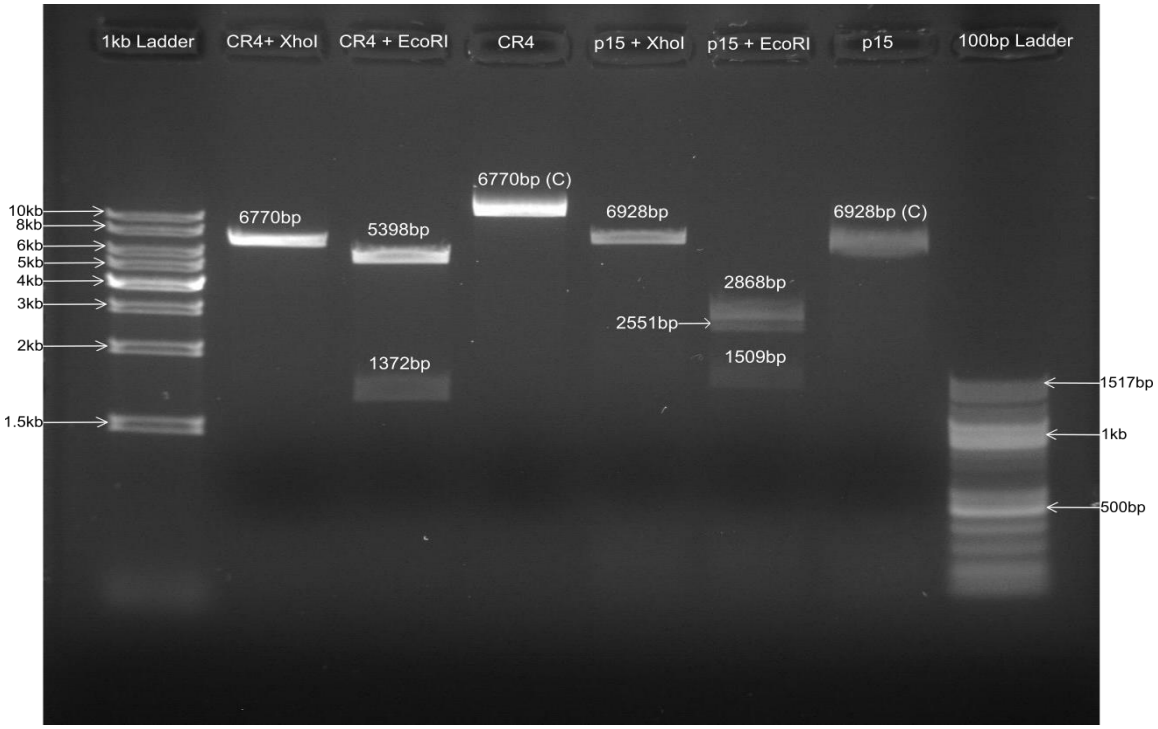

Figure 3.1. Digital image of 1% agarose gel electrophoresis of restriction digests. Results include digestion of plasmids pCas9-CR4 (CR4) and pKDsgRNA-p15 (p15) by restriction enzymes XhoI and EcoRI. Uncut, circular versions of both plasmids were also run as controls (C). Ladders used were NEB Quick-Load® Purple 1 kb Plus DNA Ladder (1kb Ladder) and NEB Quick-Load® 100 bp DNA Ladder (100bp ladder).

3.3.4 - Design of proof-of-concept deletion

With the structure of the no-SCAR plasmids confirmed, I was ready to produce a proof-of-concept deletion to establish the no-SCAR system would function as desired in our laboratory. I decided to reproduce the *ack* gene deletion carried out by Reisch & Prather (2015) which I could validate by colony PCR and Sanger sequencing. Reproducing a deletion previously created would allow for simple comparison of my results to those of Reisch & Prather (2015), quickly providing the means to assess the success of my attempt to implement the no-SCAR system. To produce the *ack* deletion, I designed: an oligonucleotide for homologous recombination around the deletion site, and primers for CPEC, colony PCR, sgRNA retargeting, PCR, and Sanger sequencing. Using the Benchling software (https://www.benchling.com/) to carry out my DNA design, with FASTA sequence for *E. coli* K-12 MG1655 retrieved from NCBI (https://www.ncbi.nlm.nih.gov/nuccore/545778205) and using Sigma-

Aldrich (https://www.sigmaaldrich.com/catalog/product/sigma/oligo?lang=en&region=GB) to order the primers and oligonucleotides, I produced the following primers and oligonucleotides: gamR, CPEC2F, pKDseq5, sgrnaR, sgrnaA, ackCD, sgRNA-ack-F, sgRNA-ack-R, ackF-PCR, ackR-PCR, and ackF-seq. Table 3.2 summarises the sequences and intended uses of the synthesised DNA. The processes in which these primers and oligonucleotides are used are outlined below. My DNA design was performed in accordance with the no-SCAR protocol published by Reisch & Prather (2017).

3.3.5 - *ack* gene deletion

I deemed the replication of the 1095bp deletion of the *ack* gene performed by Reisch & Prather (2015) to be a useful method of confirming I had successfully implemented the no-SCAR system. By having results which I could directly compare my deletion to, I would accelerate the process of troubleshooting and fine-tuning to efficiently set up no-SCAR for subsequent recombineering. The no-SCAR system has been found to be very efficient when producing gene deletions of several hundred base pairs, with efficiency reducing but still possible with deletions up to 45kbp (Reisch & Prather, 2015). The *ack* gene was therefore an ideal candidate for a proof-of-concept deletion. As the *ack* gene is located on replichore 2 of *E. coli* and its coding sequence is on the negative strand of DNA, I had to produce an oligonucleotide which had the same sequence as the positive strand sequence of *E. coli*, ensuring that the lagging strand of genomic DNA was targeted when performing recombineering. I designed the 71bp ackCD oligonucleotide identical to the published sequence (Reisch & Prather, 2015) (Table 3.2) to target the *ack* gene for deletion, while locating the mutation site within 15bp of a PAM site (2413504bp – 2413506bp CCN on the forward strand corresponding to NGG on the reverse strand) to disrupt Cas9 binding, thus facilitating Cas9-mediated counterselection of mutants. I checked the secondary structure of the oligonucleotide with mfold (http://www.unafold.org/mfold/applications/rna-folding-form.php) using default parameters, confirming the delta G was higher than -12.5kcal/mol, then added 2-4 phosphorothioate bonds at the 5’ end of the oligonucleotide to resist degradation.

3.3.6 - sgRNAs to target *ack*

Having identified a PAM site which could be used to facilitate Cas9-mediated counterselection of cells, I had to determine Cas9 target specificity by modifying the 20bp sgRNA sequence on the plasmid pKDsgRNA-ack. This is achieved by CPEC using target-specific primers which program the pKDsgRNA-ack plasmid to target different genomic loci. As the PAM site identified was on the reverse strand of genomic DNA (section 3.3.5), it was necessary to copy the reverse complement of the 20bp following the PAM on the forward strand, as this sequence was the target site. Taking this sequence and adding gttttagagctagaaatagcaag to the 3’ end of it resulted in the creation of sgRNA-ack-F. Taking the reverse complement of the 20bp sequence identified above and adding

Table 3.2. Primers and oligonucleotides designed for no-SCAR recombineering of *E. coli*. * indicates added 2-4 phosphorothioate bonds. All primers were constructed to have sufficient melting points, GC content, and GC clamp for successful PCR, with secondary structure checked.

Name	DNA Type	Sequence 5' – 3'	Process
gamR	Primer	tttataacctccttagagctcga	CPEC
CPEC2F	Primer	cggcgtcacactttgctat	CPEC
pKDseq5	Primer	cagtgaatgggggtaaatgg	Colony PCR
sgrnaR	Primer	gcctgcagtctagactcgag	Colony PCR
sgrnaA	Sequencing Primer	agctttcgctaaggatgattt	DNA Sequencing
ackCD	Oligonucleotide	G*T*TAGTACTGGTTCTGAACTGCGGTAGTTCTTCACTGGTTATCCCAACCAACGAAGAACTGG TTATCGC*G*C	Recombineering
sgRNA-ack-F	Target Specific Primer	ACCATTTACTGCATCGATGAgttttagagctagaaatagcaag	sgRNA retargeting
sgRNA-ack-R	Target Specific Primer	TCATCGATGCAGTAAATGGTgtgctcagtatctctatcactga	sgRNA retargeting
ackF-PCR	Primer	CATGCGCTACGCTCTATGG	PCR
ackR-PCR	Primer	CGCCTTTGCGTTCCATTGC	PCR
ackF-seq	Sequencing Primer	TTCCATACCCACTATCAGGTATCC	DNA sequencing

gtgctcagtatctctatcactga to the 3' end led to the creation of primer sgRNA-ack-R. Using the Cas-OFFinder software (http://www.rgenome.net/cas-offinder/) I confirmed the target site only targeted a single genomic locus. For "PAM Type" I selected "SPCas9: 5'-NRG", for the "Target Genome" I selected *Escherichia coli* (K-12, MG1655), into the query sequence box I entered the 20bp target identified above, and for "Mismatch Number" I selected "3". I ordered primers sgRNA-ack-F and sgRNA-ack-R from Sigma-Aldrich with standard desalting and 0.025µmol synthesis. While Reisch & Prather (2015) carried out the deletion in *E. coli* DH5α cells, it was possible to design an *ack* deletion in *E. coli* MG1655 using the same oligonucleotides and with primer design unaffected.

I also had the opportunity to observe and practise electroporation of electrocompetent *E. coli* cells to prepare myself to carry out my own transformation of electrocompetent cells. This training was provided by Beth Eldridge, University of Bristol.

3.3.7 - Covid-19 and beyond

It was after I received electroporation training that our laboratories were shut in response to the growing Covid-19 outbreak in the United Kingdom. This meant that despite having designed and ordered reagents and materials for the rest of the proof-of-concept *ack* gene deletion, I was unable to carry it out. Below I outline the steps I would have taken, had our laboratories remained open, as a record of the plans I had made, which could be helpful for future lab members.

3.3.8 - Retargeting plasmids for recombineering

I would have set up two PCR reactions of volume 50µl, the first with primers gamR and sgRNA-ack-F, with the second using primers CPEC2F and sgRNA-ack-R. In 0.2ml PCR tubes at room temperature I would combine 10µl 5X Colourless GoTaq Flexi Buffer (Promega, M890A), 4µl $MgCl_2$ solution (Promega, A351B), 1µl PCR nucleotide mix (Promega, C1141), 1µl each primer, 0.25µl GoTaq Hot Start Polymerase (Promega, M5001), 2µl plasmid pKDsgRNA-ack as DNA template, and 30.75µl dH_2O. I would then set a PCR reaction to run for an initial denaturation cycle at 95°C for 120s, then 30 cycles of 95°C for 8s, 64°C for 8s, and 72°C for 180s. I would then *DpnI* (Thermo Fisher Scientific, ER1701) digest the template, adding 10 units per 50µl to the reaction tube, incubating at 37°C for 15 minutes. Following this, I would run the PCR products on a 0.8% agarose gel (section 3.2.4.4). I would excise the bands produced at 3kb and 4kb, followed by gel purifying the PCR products using the QIAquick Gel Extraction Kit (Qiagen, 28704) and eluting in 1µl of dH_2O. I would then perform CPEC cloning by mixing both PCR products together in a single 0.2ml PCR tube as above, removing the DNA template and adding 32.75µl dH_2O instead. The PCR products serve as the DNA template and primers here. I would then set a PCR reaction to run for an initial denaturation cycle at 95°C for 120s, then 15 cycles of 95°C for 8s, 57°C for 15s, and 72°C for 180s.

3.3.9 - Transforming bacterial cells by electroporation

I would transform 5µl of the PCR reaction (section 3.3.8) into 50µl electrocompetent *E. coli* (section 3.2.3), pulsing using an electroporator with standard settings for a 1 mm cuvette of 1.8 kV, 200 Ω, and 25mF. These cells would recover in 500µl SOC medium (Merck, S1797-10X5ML) for 60 minutes at 30°C, and then I would plate 200µl of the solution on LB agar with 50 mg/L spectinomycin, incubating overnight at 30°C.

3.3.10 - Sanger sequencing sgRNAs

To confirm that the sgRNA was retargeted, I would pick an isolated colony with a sterile pipette tip and suspend in 100µl dH_2O, before transferring 10µl of the cell suspension to a 0.2ml PCR tube for PCR (section 3.3.8). Altered reagents include: 1µl primer pKDseq5, 1µl primer sgRNAR, and 22.75µl dH_2O. I would then set a PCR reaction to run for an initial denaturation cycle at 95°C for 120s, then 30 cycles of 95°C for 15s, 57°C for 15s, and 68°C for 60s. I would then Sanger sequence the PCR products using a BigDye Terminator v3.1 Cycle Sequencing Kit (ThermoFisher, 4337454). This involves combining in a lidded 0.2ml PCR tube 2µl BigDye Sequencing Reaction Mix, 1µl 5x Sequencing buffer, 3µl PCR product, 1µl sequencing primer sgrnaA, and 3µl dH_2O. Using a thermal cycler, I would initially cycle the solution at 94°C for 30s, then 30 cycles of 96°C for 10s, 50°C for 5s, and 60°C for 240s. To purify the sequencing products, I would transfer 10µl of the sequencing products to a clean 1.5ml microcentrifuge tube, before adding 10µl dH_2O, 5µl 125mM EDTA (Fisher Scientific, 10618973), and 60µl 100% ethanol. I would first vortex the microcentrifuge tube and then centrifuge at 13,000rpm for 10 minutes. I would then remove the supernatant and rinse the pellet with 100µl 75% ethanol, before again centrifuging at 13,000rpm for 10 minutes. Removing the supernatant again, I would dry the pellet in a vacuum, before resuspending in 15µl formamide loading buffer (Geneflow, B9-0028) and transferring to a microtiter plate for sequencing. Finally, I would recover the cloned *E. coli* cells in LB media with 50mg/L spectinomycin at 30°C.

3.3.11 - Recombineering

Once I had established that the sgRNA was successfully retargeted, I would implement recombineering and Cas9 counterselection in *E. coli* MG1655. I would prepare a liquid culture of *E. coli* DH5α containing plasmid pCas9-CR4 with 30µg/L chloramphenicol (section 3.2.2). I would then miniprep pCas9-CR4 using QIAprep Spin Miniprep Kit (50) (QIAGEN, 27104). Next, I would transform 5µl pCas9-CR4 into 50µl electrocompetent MG1655 *E. coli* using an electroporator with settings for a 1 mm cuvette of 2.5kV, 200 Ω, and 25mF. Transformed cells would recover in SOC for 60 minutes, before being plated on LB agar with 30µg/L chloramphenicol at 37°C. The next day I would pick a colony and prepare electrocompetent cells (section 3.2.3), before transforming cells with plasmid pKDsgRNA-ack using the same electroporator settings as for pCas9-CR4. I would recover cells again

in SOC for 60 minutes, then plate 200µl of the solution on LB agar with 30mg/L chloramphenicol and 50mg/L spectinomycin, incubating overnight at 30°C.

Once both plasmids required for recombineering and Cas9 counterselection were successfully transformed into *E. coli* MG1655, it would then have been possible to begin editing the bacteria's genome. I would achieve this by first inoculating a colony from the chloramphenicol/spectinomycin LB plate into 4ml of SOB (Fisher Scientific, 11337689) with 30mg/L chloramphenicol and 50mg/L spectinomycin, incubating at 30°C until the OD_{600} reaches 0.4. I would then induce λ Red by addition of L-arabinose (VWR, 1B1473-25G) to 0.2% and incubate for 15 minutes at 30°C. Next, I would place the culture on ice for 5 minutes to chill the cells, and then centrifuge at 3000 x g for 10 minutes to pellet the cells, before removing the supernatant. I would resuspend the cells in 1ml ice-cold dH_2O and pipette 1ml glycerol-mannitol solution (Fisher Scientific, 15351791) underneath the cells. I would centrifuge the cells at 3000 x g for 10 minutes before removing both layers of supernatant in turn. I would then finally resuspend the cells in 400µl glycerol-mannitol solution.

3.3.12 - Linear DNA transformation

To achieve homologous recombination-mediated genetic engineering of *E. coli* MG1655, I would take 50µl of the electrocompetent cells prepared (section 3.3.11) and add the ackCD oligonucleotide to a concentration of 2µM, using a nanodrop to inform DNA concentration. After mixing gently and transmitting to an ice-cold 1mm electroporation cuvette (Geneflow, E6-0062), I would transform the donor DNA into the cells by electroporation (section 3.3.9). One can provide a control here by transforming cells with an oligonucleotide which is untargeted by the sgRNA, which would not prevent Cas9 from binding and as such would result in cell death. I would then recover the cells in 1ml of SOC at room temperature, before incubating at 30°C and shaking at 250rpm.

3.3.13 - Serial dilution

Following culture growth, I would perform a serial dilution of five tenfold dilutions of the cell culture, spotting 8µl onto LB agar plates with 30mg/L chloramphenicol, 50mg/L spectinomycin, and 100µg/L anhydrotetracycline (aTc) (Stratech Scientific, C4291-APE-25mg). I would seal these plates with parafilm and incubate at 30°C. If I were to do the control experiment alongside the on-target oligonucleotide transformation, I would expect to see colonies about three orders of magnitude higher on the experimental plates.

3.3.14 - Genotyping colonies

I would genotype the experimental colonies by Sanger sequencing (section 3.3.10). I would pick an isolated colony (section 3.3.13) with a sterile pipette tip and suspend in 100µl dH_2O, before transferring 10µl of the cell suspension to a 0.2ml PCR tube for PCR (section 3.3.8). Altered reagents

include: 1µl primer ackF-PCR, 1µl primer ackR-PCR, and 22.75µl dH_2O. I would then set a PCR reaction to run for an initial denaturation cycle at 95°C for 120s, then 30 cycles of 95°C for 15s, 52°C for 15s, and 68°C for 60s. I would then Sanger sequence the PCR products (section 3.3.10). Altered reagents include: 1µl sequencing primer ackF-seq. Thermal cycling, purification and sequencing of products remain unchanged (section 3.3.10).

3.3.15 - pKDsgRNA-ack curing

From the serial dilution plate (section 3.3.13) I would prepare liquid culture with a single colony inoculated (section 3.2.2). I would then streak the cell culture on a LB agar plate with 30mg/L chloramphenicol and incubate overnight at 42°C. To test for spectinomycin resistance, I would patch isolated colonies onto both LB agar plates and LB agar with 50mg/L spectinomycin and incubate until I could see growth. Colonies which cannot grow on the LB agar with spectinomycin plate have therefore been cured of plasmid pKDsgRNA-ack, as it confers spectinomycin resistance. These colonies can then be used for transformation with a retargeted target plasmid for subsequent no-SCAR recombineering, enabling multiple genomic edits in the same cell line by repeating the processes outlined above. I deemed it possible, and of biological interest, to perform iterative cycles of deletions on the same cell lines, informed by the *in silico* predictions of the *E. coli* WCM, with the long-term goal of discovering a potentially novel minimal gene set for *E. coli*. Once established, the no-SCAR method of genome engineering would be able to rapidly produce gene deletions in a way which hopefully, when guided by the predictions of the WCM, would produce viable cells with a significant portion of their genome removed.

3.3.16 - pCas9-CR4 curing

When I had achieved my desired number of iterative genomic edits in each *E. coli* MG1655 cell line, I would remove plasmid pCas9-CR4, leaving the cells plasmid-free. I would miniprep pKDsgRNA-p15 using the QIAprep Spin Miniprep Kit (50), then transform the plasmid into the electrocompetent cells by electroporation (section 3.3.9). I would then recover the cells in 1ml SOC at 30°C for 2-3 hours. I would add 100µg/L aTc to induce *cas9* and the p15 targeting sgRNA. After incubating for two hours at 30°C, I would plate the cells on LB agar with 50mg/L spectinomycin and 100ng/L aTc, incubating again at 30°C overnight. Finally, I would screen 10-20 colonies by patching onto both LB agar and LB agar with 30mg/L chloramphenicol, incubating at 37°C to confirm the loss of chloramphenicol resistance.

3.3.17 - Post proof-of-concept deletion

Identifying gene targets to pursue beyond a proof-of-concept deletion was a very involved process that was a point of continuous discussion among members of our group. An ambitious goal was to

begin deletions in genome-reduced *E. coli* strains such as DGF-298 (Hirokawa *et* al., 2013) and MDS42 (Pósfai *et al.,* 2006), with the aim of creating further genome reduction. This would be achieved by iterative cycles of gene deletions using the no-SCAR system described above, deleting gene after gene as directed by the single gene knockout simulations performed using the *E. coli* WCM. In this way, we would successfully combine *in silico* and *in vivo* research by streamlining the process of selecting candidate genes for deletion while maintaining viable cells. With an initial five-day duration for an initial no-SCAR mutation, with a subsequent three-day turnaround for each mutation thereon, this process would still be time-consuming. Assuming perfect conditions, it would have been possible to produce ~40 deletions using the no-SCAR method in the time I had allocated myself for laboratory research. Subsequent mutations would have been beyond the scope of this project. Furthermore, when modifying *E. coli* MDS42, it is necessary to question the accuracy of the *E. coli* WCM's predictions, as the model is a hybrid of *E. coli* strains MG1655, B/r, and BW25113 (Figure 1.1). Mutations in *E. coli* MDS42 could behave very differently to the *in silico* predictions. The process of acquiring *E. coli* MDS42 also proved difficult, with the vendor I contacted responding slowly to enquiries.

3.3.18 - General troubleshooting

I had to solve various problems in the laboratory, including having LB media which was being contaminated with bacterial growth multiple times. This was despite me following aseptic technique to the best of my ability. The crucial change I made was to cease working in a laminar airflow hood, which may have had filters that needed changing amongst other problems. I instead reverted to work on the bench with a Bunsen burner, which I believe reduced the number of variables in my workspace that could lead to contamination. As explained above, I also overcame the challenges posed by the restriction digests of plasmids which allowed me to eventually confirm their structure (section 3.3.3).

3.4 - Discussion

While not being able to complete a proof-of-concept deletion due to factors beyond my control, the steps I took in implementing the no-SCAR protocol in our lab shall serve as a springboard for subsequent researchers who are interested in doing similar work. We have organised the delivery of *E. coli* MG1655 and DHα stocks containing plasmids pCas9-CR4, pKDsgRNA-ack, and pKDsgRNA-p15 which I prepared to Rayane Nunes Lima, a collaborator in Brazil, showing that the progress I made is having a useful impact. Through the construction of this thesis and the outlining of the methodology that I followed, any subsequent researchers in our lab should be able to implement the no-SCAR system more efficiently, as I attempted to make the methods as explicit and reproducible as possible. This includes processes which are alluded to vaguely by Reisch & Prather (2017), for

example “genotype the colonies…by Sanger sequencing” which fails to detail the PCR reactions required for such a process. Through trying to establish this method of genome engineering, we also put into place useful systems for future use by our research group. This includes setting up the pre-purchased RapidRED card, which permits easy ordering of synthesised oligonucleotides and primers from Merck (https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/promo_NOT_INDEXED/General_Information/1/edoc13-order-rapidred-card.pdf). If members of the lab group want to carry out genome editing as I intended, they will be able to do so at an accelerated rate.

Chapter 4 – *E. coli* WCM single gene knockout analysis

4.1 - Introduction

WCMs permit *in silico* investigation of the literature published on the organism that they simulate. The *M. genitalium* and the *E. coli* WCMs have been used for this research (Karr *et al.,* 2012; Macklin *et al.,* 2020). It is necessary to compare WCMs against data from the literature that was not used to construct them, to understand how accurately these models can predict phenotypes. The Covert Lab, Stanford, gave us pre-publication access to their *E. coli* WCM and I hoped to test the model's ability to predict the effects of creating single gene knockouts on the *in silico* cell's growth, which I would use to inform *in vivo* genome recombineering (see Chapter 3). Essentiality (see Chapter 1 section 1.4) provides a simple means of assessing the impact of single gene knockout studies on a cell: if the cell can undergo division when a gene is deleted, the gene is "non-essential" (Rancati *et al.,* 2018). If cell division is prevented, the gene is instead "essential".

Once the *E. coli* WCM was successfully adapted for use on the BlueCrystal supercomputer cluster (see Chapter 3 section 3.1), a process which took several months, we were presented with the first chance to test WCM results *in vivo*. The Keio collection (Baba *et al*., 2006) is a single gene knockout library created by targeting 4288 *E. coli* K-12 genes with single gene knockouts. Of the 4288 targeted genes, 3985 mutants were obtained, indicating the majority of genes had non-essential phenotypes when knocked out. Most of the 303 genes which did not produce mutants when deleted are candidates for essential genes (Baba *et al.,* 2006). Baba *et al.* (2006) produced essentiality scores for the 4288 genes investigated, combining their assessment of essentiality with those of previous researchers. As the output of the *E. coli* WCM produces data on growth and cell division, I would be able to assign essentiality labels to each gene deleted *in silico* and compare them with the Keio collection to assess whether the *E. coli* WCM could accurately predict the effect of removing both non-essential and essential genes from the *E. coli* genome. To organise and conduct simulations, the *E. coli* WCM uses FireWorks. This open-source tool was designed for managing high-throughput workflows using supercomputers (Jain *et al.,* 2015). FireWorks uses an online MongoDB database to store workflows which can be retrieved by a supercomputer, executed, and have the results uploaded to the online database. FireWorks allows for workflows to be run in parallel on multiple supercomputers and for failures to be automatically rectified, increasing the efficiency with which simulations can be conducted.

4.2 - Methods

4.2.1 - Data sources

I downloaded the *E. coli* WCM from GitHub (https://github.com/CovertLab/WholeCellEcoliRelease). I followed installation instructions for the *E. coli* WCM created by Dianaimh Greene, of the Advanced Computing Research Centre, and Joshua Rees-Garbutt (https://github.com/squishybinary/Ecoli_whole-cell_model_analysis). I adapted the script FireWorksBox_5.sh produced by Joshua Rees-Garbutt (https://github.com/squishybinary/Ecoli_whole-cell_model_FireWorksBoxes/blob/master/FireWorksBox_5.sh). I accessed the essentiality scores for the Keio collection published by Baba *et al.* (2006) at (https://www.embopress.org/doi/full/10.1038/msb4100050) – Supplementary Table 3 contained data for Keio collection deletion mutants.

4.2.2 - Procedures

4.2.2.1 - Setting up the *E. coli* WCM

Having downloaded and installed the *E. coli* WCM on the University's BlueCrystal supercomputer (section 4.2.1), I used the WinSCP client (https://winscp.net/eng/index.php) for File Transfer Protocol (FTP) and the PuTTY client (https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html) for Secure Shell (SSH) interfacing with the supercomputer.

4.2.2.2 - Using FireWorks to run the *E. coli* WCM

I followed a methodology outlined by Joshua Rees-Garbutt (https://research-information.bris.ac.uk/ws/portalfiles/portal/254360018/Joshua_Rees_Garbutt_Thesis.pdf) to run the FireWorks workflow tool, allowing multiple supercomputers to work in parallel and execute jobs rapidly. I created three .yaml files and uploaded them via FTP to /newhome/jr16196/wholecell2/wcEcoli/wholecell/fireworks using the file my_fworker.yaml. A second file, my_launchpad.yaml, was changed to access my desired online database and read:

```
authsource: heroku qc6nn407
host: ds345028.mlab.com
logdir: null
name: heroku_qc6nn407
password: ************
port: 45028
ssl: false
ssl_ca_certs: null
ssl_certfile: null
ssl_keyfile: null
ssl_pem_passphrase: null
```

```
strm_lvl: INFO
user_indices: []
username: jr16196
wf_user_indices: []
```

The third file, my_qadapter.yaml, was adapted to correspond with my supercomputer account and directory, to read:

```
logdir:
/newhome/jr16196/wholecell2/wcEcoli/wholecell/fireworks/logs/qadapte
r
_fw_name: CommonAdapter
_fw_q_type: PBS
rocket_launch: rlaunch -w
/newhome/jr16196/wholecell2/wcEcoli/wholecell/fireworks/my_fworker.y
aml -l
/newhome/jr16196/wholecell2/wcEcoli/wholecell/fireworks/my_launchpad
.yaml singleshot
nnodes: 1
ppnode: 1
walltime: '10:00:00'
account: 'jr16196'
job_name: 'jr16196ecoli'
email: 'jr16196@bristol.ac.uk'
notification_options: 'bea'
pre_rocket: null
post_rocket: null
```

4.2.2.3 - Using FireWorks to create single gene knockout simulations

I created the bash script FireWorksBox_3.sh (https://github.com/JJRightmyer/SingleGeneKO/blob/main/FireWorksBox_3.sh) and uploaded it via FTP to /newhome/jr16196/wholecell2/wcEcoli/wholecell/fireworks. This script provides a description of the simulation to the script fw_queue.py which translates it into a workflow. My FireWorksBox_3.sh script was adapted from: https://github.com/squishybinary/Ecoli_whole-cell_model_FireWorksBoxes/blob/master/FireWorksBox_5.sh. FireWorksBox_3.sh was created to produce 1214 simulations, each knocking out a single gene and running the resulting model for two generations. I then created a directory in /newhome/jr16196/wholecell2/wcEcoli/wholecell/fireworks, called /logs, in which I created two subdirectories called /launchpad and /qadapter. To set the supercomputer to fetch and launch FireWorks I ran the following command:

```
qlaunch -r -l my_launchpad.yaml -w my_fworker.yaml -q my-
qadapter.yaml rapidfire –nlaunches infinite –sleep 30 –maxjobs_queue
100
```

I then made FireWorksBox_3.sh executable and ran it using commands:

```
cd /newhome/jr16196/wholecell2/wcEcoli/wholecell/fireworks

chmod u+x FireWorksBox_3.sh

./FireWorksBox_3.sh
```

From the simulation output, the massFractionSummary.png and histogramDoublingTime.png files were of interest as they provided graphs of changing cellular components and cell division respectively as the simulation proceeded. I manually analysed the massFractionSummary.png graphs to assess the essentiality of each knocked-out gene in the simulation.

4.2.2.4 - Using FireWorks to create wild-type simulations

I altered FireWorksBox_3.sh to produce a script which described 30 wild-type simulations, which were run in the *E. coli* WCM for two generations. I named this script FireWorksBox_4.sh (https://github.com/JJRightmyer/SingleGeneKO/blob/main/FireWorksBox_4.sh) and it was run using the same methodology as outlined in Section 4.2.2.3. Due to time constraints, Ioana Gherman kindly ran these simulations for me on BlueCrystal.

4.2.2.5 - Keio collection

I downloaded Supplementary Table 3 from Baba *et al.* (2006) (section 4.2.1) and manually compared their assessment of essentiality against the output of the *E. coli* WCM.

4.2.3 - Data storage

The data I produced from two runs of 1214 single gene knockout simulations using the *E. coli* WCM can be found at the University of Bristol Advanced Computing Research Centre Research Data Storage Facility (https://data.bris.ac.uk/data/dataset/1v58mpokwp9927bf6s36osb5k). The growth data for 30 wild-type simulations performed by me and Ioana Gherman can also be found here (https://data.bris.ac.uk/data/dataset/2uymxzuikvnc526gu3b44gvxht).

4.3 - Results

4.3.1 - Issues with running gene knockout simulations

An initial run of 1214 single gene knockout simulations produced massFractionSummary.png files with inaccurate labels (Figure 4.1). The GeneID label attributed to the graphed data was different to the GeneID in the graph's title, with the mismatch being inconsistent across the simulations (between 2-4 genes off each time). Our team thought that the issue, despite requiring manual fixing due to its inconsistent nature, was superficial. We thought that we had still knocked out our intended gene targets. I discovered this was not the case and we had unintentionally knocked out

360 unmodelled genes. Joshua Rees-Garbutt identified a mistake within the rnas.tsv file in directory ./reconstruction/ecoli/flat which caused the first entry (EG30099) in rnas.tsv to be ignored, shifting the gene index down by one. This caused all subsequent gene knockouts to be inaccurate. Joshua Rees-Garbutt proceeded to fix the mistake in the rnas.tsv file which caused the simulations to knock out unmodelled genes and my next run of simulations was successful. This took two weeks.

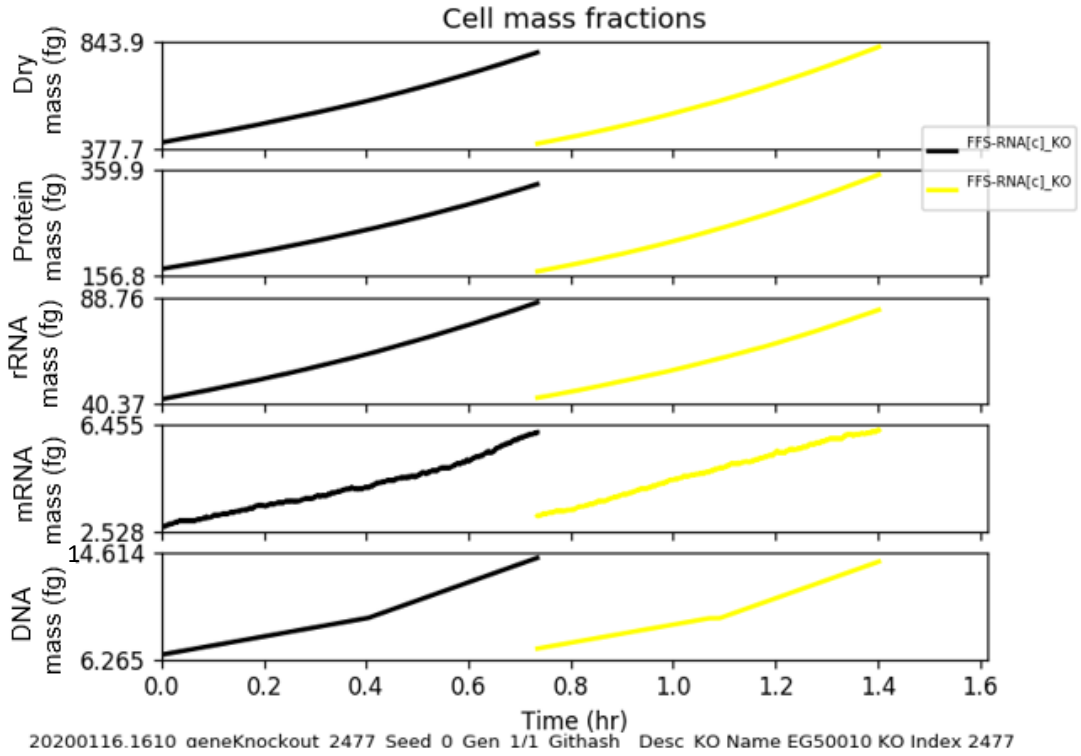

Figure 4.1. The graphed output of a single gene knockout using the *E. coli* WCM. Each line represents a single generation of the cell growing. Changes in the cell's mass, protein, rRNA, mRNA, and DNA are plotted over time. This simulation was supposed to knock out *frmA (*EG50010) but instead knocked out the unmodelled gene *ffs* (FFS-RNA[c]_KO). The first generation of the cell is shown in black, and the second in yellow.

4.3.2 - 1214 single gene knockouts

Having set up the *E. coli* WCM (Section 4.2.2.1), I conducted the second set of 1214 single gene knockout simulations for two generations. I used two generations due to computational space restrictions while requiring multiple generations to assess gene essentiality. Of the 1214 modelled genes which were knocked out in individual *in silico* cells, 899 genomes successfully produced output of massFractionSummary.png, histogramDoublingTime.png, and short_name files. The file named massFractionSummary.png details growth metrics for the simulations including: dry mass, protein mass, rRNA mass, mRNA mass, and DNA mass (Figure 4.1). The histogramDoublingTime.png file shows if the *in* silico cell has divided, which is also shown by massFractionSummary.png. The

short_name file details which gene was knocked out in the model. The genes that were knocked out in these 899 genomes were non-essential, as cell division occurred. 313 genes produced no output when knocked out, as *in silico* cells did not divide and therefore produced no automated graphs. This could imply that these genes were essential or could mean there was a problem with the code when producing these knockouts. Genes *gltX* (EG10407) and *gor* (EG10412) produced histogramDoublingTime.png and short_name files when knocked out but failed to produce a massFractionSummary.png file. This was due to a computational error and while the *in silico* cells did divide, I could not analyse the cellular components which underlay the cell's behaviour. I decided against their inclusion in my analysis of the simulations.

4.3.3 - Keio collection issues

Comparing the results of the single gene knockouts produced by the *E. coli* WCM to the essentiality score given by Baba *et al.* (2006) in the Keio collection demanded an unexpected step in data analysis. Supplementary Table 3 (https://www.embopress.org/doi/full/10.1038/msb4100050) contained the raw data Baba *et al.* (2006) used to define gene essentiality, which they accomplished by calculating a score based on essentiality data from Gerdes *et al.* (2003), Hashimoto *et al.* (2005), Kang *et al.* (2004), and their own data. This was done to allow comparison with other *E. coli* gene essentiality studies. The only way to separate the data produced by Baba *et al.* (2006) from the essentiality score they provide was to manually calculate and deduct the scores attributed to essentiality data from Gerdes *et al.* (2003), Hashimoto *et al.* (2005), and Kang *et al.* (2004). I chose to separate the data from Baba *et al.* (2006) from those of the other researchers as Baba *et al.* (2006) were the only group to publish results of a single gene knockout experiment. The other researchers employed strategies such as genetic footprinting and transposon mutagenesis, to which "striking differences" could be attributed (Baba *et al.,* 2006). I believed a direct comparison between my own single gene knockout experiment and one found in the literature would produce the most interesting result. Once I found the score Baba *et al.* (2006) gave their own data, I could assign the gene essentiality result they discovered from their *in vivo* experiments. This was a time consuming process which highlighted the importance of clear data presentation to me. The difficulty I experienced was exacerbated by the website used by Baba *et al.* (2006) to distribute the Keio collection (http://ecoli.aist-nara.ac.jp/), which was down for the duration of my project.

4.3.4 - Comparing Keio collection essentiality to model output

Once I had interpreted the *in vivo* gene essentiality result from Baba *et al.* (2006), I was able to compare this to the results of the single gene knockout experiment outlined in Section 4.3.2. I manually constructed a table to compare the model output to the Keio collection, first inputting the Keio collection assessment of essentiality, then comparing that to the *E.* coli WCM output and

assigning a key to document the agreement or disagreement of the two (Appendix 17). A challenge of this was reconciling the gene codes the Keio collection and the *E. coli* WCM had used, as they were different and required manual fixing (e.g. the *alr* gene had an ID of EG10001 in the model but ECK4045 in the Keio collection). I then created a summary table for the results of this comparison (Table 4.1). Of the 1212 modelled genes I analysed, 834 agreed with the Keio collection, with 727 non-essential genes producing graphs of standard growth and 64 essential genes not producing output, and 43 non-essential genes producing graphs of impacted growth.

Table 4.1. A summary of the *E. coli* whole-cell model (WCM) single gene knockouts output when compared to essentiality data published in the Keio collection. A breakdown of each result is provided with reasons for agreement and disagreement. In bold are the total numbers of genes in the *E.* coli WCM which agree and disagree with the Keio collection, alongside genes whose essentiality could not be assessed.

Gene phenotype in the Keio collection	Phenotype in the *E. coli* WCM	Agreement between the Keio collection and the *E. coli* WCM	Genes
Non-essential	Standard growth	Agree	727
Non-essential	Impacted growth	Agree	43
Essential	No output	Agree	64
			834
Essential	Standard growth	Disagree	115
Essential	Impacted growth	Disagree	7
Non-essential	No output	Disagree	249
			371
Uncertain essentiality	Standard growth	Uncertain	**7**

30 wild-type simulations (section 4.2.2.4) provided values of growth for cells showing a standard growth phenotype (Table 4.2). If any of the cell's dry, protein, rRNA, mRNA or DNA mass values were lower than the final growth range values at the end of the second generation of a simulated cell, I deemed the knockout to have produced the impacted growth phenotype. Of the 371 genes which disagreed with the Keio collection when simulated, 115 were genes that Baba *et al.* (2006) showed were essential *in vivo,* which produced graphs of standard growth when knocked out in the model, seven were essential genes which produced graphs of impacted growth when knocked out, and 249 were non-essential genes which did not produce output. Seven genes also produced graphs of normal growth but were given uncertain essentiality by Baba *et al.* (2006), so they could neither agree nor disagree with the simulation results. This highlights that the *E. coli* WCM currently has no designated output for *in silico* cells which fail to divide. Instead, we assumed that because the

knocked out gene caused the model to produce no output, it may be essential to the cell's growth. This is now being investigated more carefully by other members of our group.

Table 4.2. The ranges for Y-axis values of 30 wild-type simulations at the end of two generations of growth using the *E. coli* WCM.

	Cell mass ranges	Protein mass ranges	rRNA mass ranges	mRNA mass ranges	DNA mass ranges
Final growth values	652.1 - 896.6	301.9 - 398.4	80.64 - 106.07	3.808 - 7.421	11.098 - 15.151

Of the single gene knockouts described above, seven (EG10088, EG10205, EG10206, EG10207, EG10208, EG11314, EG11539) were essential genes which successfully produced output, indicating they were non-essential to the *in silico* cell's growth and disagreeing with the Keio collection. The massFractionSummary.png files produced by these simulations were impacted compared to the standard output graphs, with dry mass, protein mass, rRNA mass, mRNA mass, and DNA mass lower than the ranges which indicate the standard growth phenotype (Figure 4.2). This was also the case for 43 non-essential single gene knockouts which agreed with the Keio collection, with massFractionSummary.png files different to the standard output graphs (Figure 4.2). The inconsistent axes of the graphs produced by the *E. coli* WCM made visual interpretation and comparison of the outputs difficult (https://data.bris.ac.uk/data/dataset/1v58mpokwp9927bf6s36osb5k). As these graphs are produced automatically by the model, modifying them requires familiarity with details of the model, which are time-consuming to obtain. The graphs I present in Figure 4.2 were manually reconstructed to allow comparison between the model outputs. Nevertheless, the deviation of these cells from the standard growth pattern warrants further investigation, for example repeating these single gene knockouts over more generations to see if cells will continue to divide beyond two generations. This may resolve the issue as the model's creators suggested that initial metabolite pools may sustain the cell for two generations even when an essential gene is removed (*personal correspondence)*.

It may be possible to investigate why 313 single gene knockouts produced no output when run using the *E. coli* WCM through gene knockdown experiments. Ioana Gherman of our research group has since shown that creating 90% knockdowns allows most of the simulated cells to divide and is continuing to investigate 97% and 100% knockdowns. As high percentage gene knockdowns should produce cells with similar phenotypes to gene knockouts, this should reveal if there is an issue with the way the model is built that causes knockouts to fail for spurious reasons.

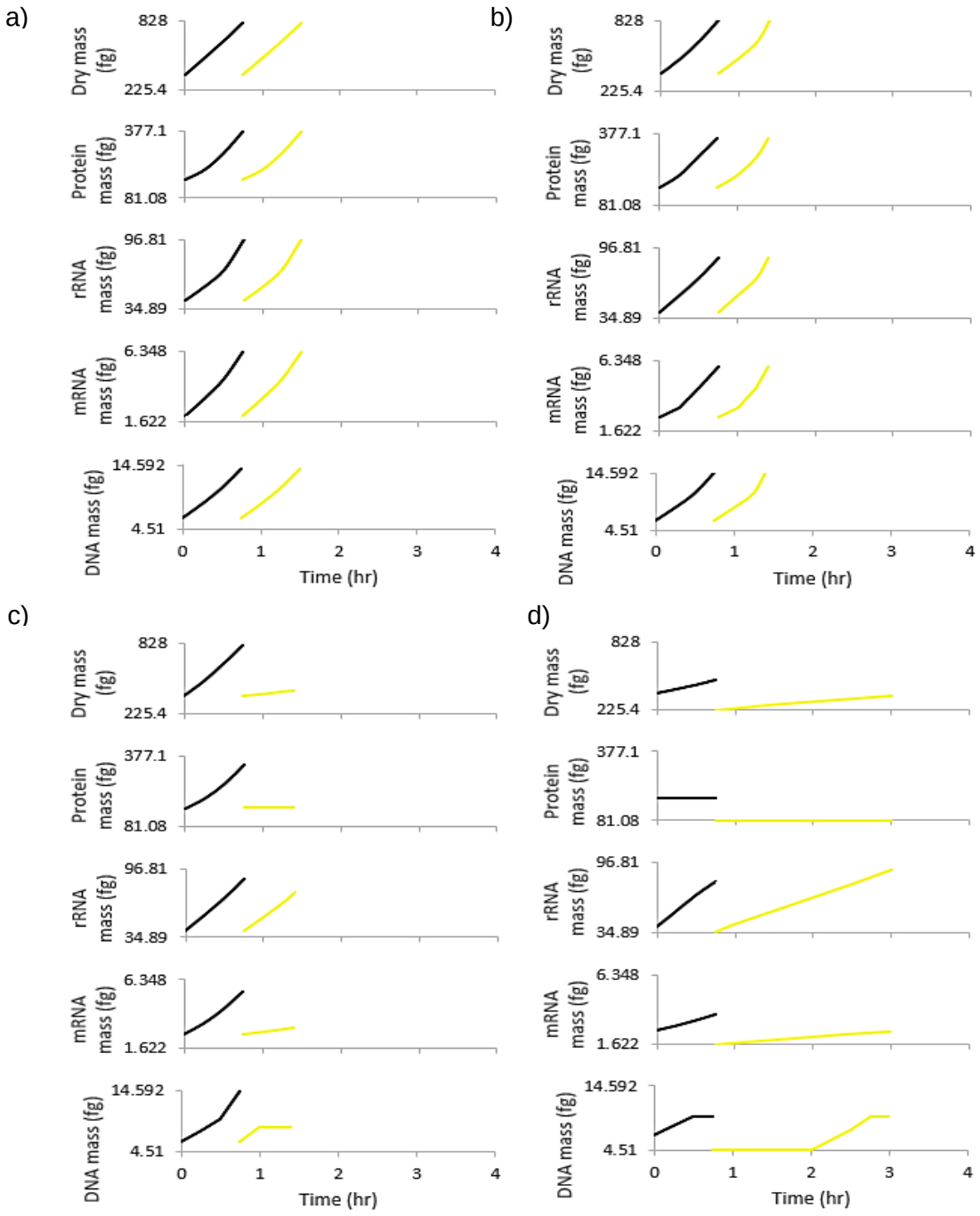

Figure 4.2. Plots of four *E.* coli WCM simulation outputs for cells grown for two generations with axes adjusted to enable comparison. Dry, protein, rRNA, mRNA, and DNA mass of cells were plotted over time. The first generation is plotted in black and the second in yellow. a) a wild-type simulation from which the standard growth phenotype was established. b) a single gene knockout simulation of *alr* (EG10001) showing the standard growth phenotype. c) a single gene knockout simulation of *aroE* (EG10077) showing the impacted growth phenotype in the second generation. d) a single gene knockout simulation of *asd* (EG10088) showing the impacted growth phenotype in both generations.

4.4 - Discussion

This work represents the first study of gene essentiality that combines the *E. coli* WCM and the literature. Comparing the 1214 modelled genes in the *E. coli* WCM to the Keio collection, a single gene deletion library deemed the gold standard for identifying essential genes (Goodall *et al.,* 2018), has allowed us to assess the model's predictive capabilities when creating single gene knockouts. My preliminary results suggest the *E. coli* WCM produces single gene knockouts which agree with at least 68.8% of the Keio collection over two generations. This figure is promising, but there is also much room for improvement. Conversations with the model's creators have revealed that computational bugs could have caused 313 single gene knockouts to produce no output. This is not unexpected as they have not yet explored how the model responds to gene knockouts. The results of Ioana Gherman's 90% knockdown experiment suggest that there are problems in the model when responding to gene knockouts, as opposed to knockdowns. This probably indicates that my results are caused by problems with the model's code not working the way it was intended when producing gene knockouts, which I have relayed to the model's creators. This has already led to better procedures for modelling gene function modifications in the *E. coli* WCM, which will make it much more useful for future research that uses the model to investigate the effects of modifying gene functions.

My results also establish gene targets for no-SCAR recombineering (see Chapter 3) where there is both agreement and disagreement between the *E. coli* WCM output and the Keio collection. Seven genes (EG10052, EG10053, EG10055, EG10527, EG11006, EG11866, EG11868) which were not required for cell division in my simulations, were labelled as possessing uncertain essentiality (Baba *et al.,* 2006) and therefore are strong candidates for investigation using the no-SCAR method. This research could help provide more data regarding the essentiality of these genes, for example *araA* (EG10052) and *rhaD* (EG11866) whose essentialities have proved inconclusive (Gerdes *et al.,* 2003; Baba *et al.,* 2006; Goodall *et al.,* 2018). The no-SCAR method also permits the simple creation of multiple gene deletions in a cell, which would aid our understanding of possible effects of genomic context on the essentiality of these genes (Rancati *et al.,* 2018). As a cell's need for specific genes changes with the presence or absence of other genes in its genome, it could be the case that these genes of uncertain essentiality are dependent on their genomic context (Rees-Garbutt *et al.,* 2020c). Further *in vivo* investigation of the *E. coli* WCM's predictions, accurate or otherwise, will help establish how useful this model can be when designing genetic edits in *E. coli*. This work, alongside repeating the single gene knockouts with more generations, repeating the simulations to test for consistency between the simulations, and potentially performing multiplex gene knockout *in silico* will give further indication of the *E. coli* WCM's ability to usefully inform *E. coli* genome research.

Chapter 5 – Improving the *E. coli* WCM

5.1 - Introduction

Despite severely impacting my project's progress in the laboratory, the Covid-19 pandemic sparked further collaboration between our research group and that of Markus Covert at Stanford University, who created both the *E. coli* and *M. genitalium* WCMs which have served as focal points of my research. The collaboration was designed to develop the *E. coli* WCM beyond its release snapshot, utilising the skillsets of researchers outside Stanford to help propel the *E. coli* WCM towards being a true WCM, able to account for the function of all well-annotated *E. coli* genes. This would create a model which shows greater fidelity to *in vivo E. coli*. This would permit cross-evaluation of *E. coli* datasets in the literature beyond what the Covert Lab has already achieved (Macklin *et al.,* 2020) and provide an increasingly useful model for genome design *in silico*.

Initial communication with the Covert Lab was focused on deciding which elements of the *E. coli* WCM could be usefully improved by researchers who were relatively unfamiliar with the inner workings of the model. Conversations with Markus Covert, Travis Horst, Gwanggu Sun, Joshua Rees-Garbutt, and Sophie Landon led me to begin investigating how cellular metabolism was implemented in the *E. coli* WCM. The metabolic model of the *E. coli* WCM has detailed quantitative parameters for 380 metabolic reactions (Macklin *et al.,* 2020) and is a reconstruction of the *E. coli* K-12 model published by Weaver *et al.* (2014), which was in turn based on the Orth *et al.* (2011) metabolic model of *E. coli*. The metabolic model is also able to adjust to time-dependent/cell cycle-dependent behaviour from other simulated processes while maintaining homeostasis, representing a significant improvement over the *M. genitalium* WCM (Karr *et al.,* 2012). Macklin *et al.* (2020) incorporate kinetics-based and FBA-based approaches to make use of the many reported parameter values for *E. coli* without additional model fitting to ensure metabolism is readily integrable with the rest of the WCM. However, there is scope to add parameters for many more metabolic reactions, with 1913 metabolic reactions published to the EcoCyc database for *E. coli* MG1655 (Keseler *et al.,* 2017). My goal was to successfully implement metabolic reactions which I found interesting and could direct the development of the *E. coli* WCM towards unique applications.

I also worked with Travis Horst to improve transcriptional regulation within the *E. coli* WCM. The quantitative model of transcriptional regulation in the *E. coli* WCM incorporates the function of 22 transcription factors regulating 355 genes, in turn describing 438 regulatory interactions (Macklin *et al.,* 2020). Implementing transcriptional regulation requires modelling the activation or inhibition of a transcription factor and once it is active, modelling its effect on RNA polymerase recruitment to a promoter site. Macklin *et al.* (2020) separated transcription factors into three classes based on their

mechanism of activation. The first, one-component systems, are transcription factors that are directly activated or inhibited by a small molecule ligand. The second, two-component systems, are transcription factors that are paired with a separate sensing protein that responds to an environmental stimulus. The sensing protein phosphorylates the cognate transcription factor in this case. The third, zero-component systems, are transcription factors that are active whenever they are expressed. This includes transcription factors modulated by complex feedback loops which are simplified in the model to be always active to compensate for lack of understanding. There is capacity for implementing further transcriptional regulation in the *E. coli* WCM, as the EcoCyc database details 204 transcription factors describing 6399 regulatory interactions (Keseler *et al.,* 2017). My goal was to enhance the representation of transcriptional regulation within the *E. coli* WCM to improve the model's ability to respond to a range of environmental conditions.

5.2 - Methods

5.2.1 - Data sources

I downloaded the version of the *E. coli* WCM under development from the Covert Lab's private GitHub repository (https://github.com/CovertLab/wcEcoli). I used the EcoCyc database for analysing metabolic reactions (https://ecocyc.org/). I obtained gene regulation data for Lrp from Supplemental file 4 published by Kroner *et al.* (2019) (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6349092/).

5.2.2 - Procedures

5.2.2.1 - Setting up the *E. coli* WCM

I followed installation instructions created by Sophie Landon to install and run the current version of the *E. coli* WCM on BlueCrystal. I used the PuTTy software to run the following code:

```
module add tools/git-2.22.0

#Create a directory called wholecell3

mkdir wholecell3
cd wholecell3/

git clone https://github.com/CovertLab/wcEcoli
cd wholecell3/

git clone https://github.com/pyenv/pyenv.git ~/.pyenv
echo 'export PYENV_ROOT="$HOME/.pyenv"' >> ~/.bash_profile
echo 'export PATH="$PYENV_ROOT/bin:$PATH"' >> ~/.bash_profile
echo -e 'if command -v pyenv 1>/dev/null 2>&1; then\n eval "$(pyenv
init -)"\nfi' >> ~/.bash_profile
source ~/.bash_profile

git clone https://github.com/pyenv/pyenv-virtualenv.git $(pyenv
root)/plugins/pyenv-virtualenv
```

```
echo 'eval "$(pyenv virtualenv-init -)"' >> ~/.bash_profile
git clone https://github.com/pyenv/pyenv-virtualenvwrapper.git
$(pyenv root)/plugins/pyenv-virtualenvwrapper
source ~/.bash_profile

PYTHON_CONFIGURE_OPTS="--enable-shared" pyenv install 3.6.12

cd wcEcoli/
pyenv local 3.6.12
pyenv virtualenv wcEcoli3
pyenv local wcEcoli3

pip install --upgrade pip setuptools virtualenv virtualenvwrapper
virtualenv-clone wheel

cd ../

git clone https://github.com/xianyi/OpenBLAS
cd OpenBLAS
git checkout v0.3.7

#Set Environment:

export FC=/cm/shared/languages/GCC-7.1.0/bin/gfortran
export CC=/cm/shared/languages/GCC-7.1.0/bin/gcc
FC=/cm/shared/languages/GCC-7.1.0/bin/gfortran make -j 8

make PREFIX=/newhome/jr16196/wholecell3/install install

>>Create numpy-site.cfg:

~/.numpy-site.cfg

and include:

[openblas]
libraries = openblas
library_dirs = /newhome/jr16196/Software/wholecell3/install/lib
include_dirs = /newhome/jr16196/Software/wholecell3/install/include
runtime_library_dirs =
/newhome/jr16196/Software/wholecell3/install/lib

cd ../wcEcoli/

##Install in this order##

pip install scipy==1.5.2 --no-binary --force-reinstall
pip install numpy==1.19.2 --no-binary --force-reinstall

CPATH=/usr/include/glpk CVXOPT_BUILD_GLPK=1 pip install -r
requirements.txt --no-binary scipy,cvxopt
export PYTHONPATH="$PWD:$PYTHONPATH"

make clean compile

##Test:
```

```
python runscripts/manual/runParca.py
```

```
python runscripts/manual/runParca.py --cpus 8
```

5.2.2.2 - Metabolism

In the file wcEcoli/reconstruction/ecoli/dataclasses/process/metabolism.py (https://github.com/JJRightmyer/PrivateWCMEcoli/blob/main/metabolism.py) I changed line 41 from:

```
VERBOSE = False
```

To

```
VERBOSE = True
```

This would create an output from the parameter calculator (ParCa) of all successful and unsuccessful metabolic reactions, which I could then analyse.

I then ran the parameter ParCa with the following command to reveal any metabolic reactions which could be implemented:

```
python runscripts/manual/runParca.py
```

I copied the ParCa output into a file I named Parca_output.txt (https://github.com/JJRightmyer/AddingToWCMEcoli/blob/main/Parca_output.txt) and identified the missing metabolic reactions, copying them into a file called Metabolism_targets.txt (https://github.com/JJRightmyer/AddingToWCMEcoli/blob/main/Metabolism_targets.txt).

To implement the thiosulfate sulfurtransferase reaction (THIOSULFATE-SULFURTRANSFERASE-RXN in reactions.tsv) I used the EcoCyc database to find a stoichiometrically balanced reaction (https://ecocyc.org/ECOLI/NEW-IMAGE?type=REACTION&object=THIOSULFATE-SULFURTRANSFERASE-RXN). I then added this reaction to the file wcEcoli/reconstruction/ecoli/flat/reactions.tsv (https://github.com/JJRightmyer/PrivateWCMEcoli/blob/main/reactions.tsv) to read:

```
"THIOSULFATE-SULFURTRANSFERASE-RXN"   {"S2O3[p]": -1, "HCN[p]": -1, "HSCN[p]": 1, "SO3[p]": 1, "PROTON[p]": 2}  false ["CPLX0-8219", "EG11600-MONOMER", "EG10780-MONOMER", "CPLX0-242"]
```

This reaction indicates that $S_2O_3 + HCN \rightarrow HSCN + SO_3 + 2H^+$, with all reactants and products occurring in the periplasm ([p]) in a non-reversible reaction (false) and being catalysed by four enzymes (CPLX0-8219, EG11600-MONOMER, EG10780-MONOMER, CPLX0-242).

I then uploaded the updated reactions.tsv file to BlueCrystal and ran the ParCa again using the following command:

```
python runscripts/manual/runParca.py
```

I copied the output from the PuTTy command console into a file called Thiosulfate_implemented.txt (https://github.com/JJRightmyer/AddingToWCMEcoli/blob/main/Thiosulfate_implemented.txt) where line 71 indicates successful implementation of the thiosulfate transferase reaction. I sent my updated wcEcoli/reconstruction/ecoli/flat/reactions.tsv file to Travis Horst for inclusion in the *E. coli* WCM.

5.2.2.3 - Transcriptional regulation

I downloaded fold change data for the transcription factor Lrp from Supplemental file 4 published by Kroner *et al.* (2019). I calculated the mean fold change for each gene from the log2_expr_ratio across the Log, Trans, and Stat sheets on minimal media (MIN). I calculated standard deviation for each value. I assigned a reaction direction based on whether the average expression ratio was positive or negative (either 1 or -1 respectively). I inputted this data into a file called Lrp_fold_changes.tsv (https://github.com/JJRightmyer/AddingToWCMEcoli/blob/main/lrp_fold_changes.tsv), ensuring it matched the formatting of the foldChanges.tsv file in the model and sent it to Travis Horst for inclusion in the *E. coli* WCM.

5.2.3 - Data availability

The data I produced while implementing the thiosulfate sulfurtransferase reaction and adding the transcriptional regulation by Lrp are available on GitHub (https://github.com/JJRightmyer/AddingToWCMEcoli). Where I have used or edited files which appear in the Covert Lab's private GitHub repository for the *E. coli* WCM, I have uploaded these files to a private GitHub repository of my own so I can share my work whilst protecting their work (https://github.com/JJRightmyer/PrivateWCMEcoli).

5.3 - Results

5.3.1 - Improving the metabolic model in the *E. coli* WCM

I successfully implemented the thiosulfate sulfurtransferase reaction into the *E. coli* WCM. This reaction is a pathway for the metabolism of cyanide, which has been observed in *E. coli* isolated from gold extraction circuit liquids (Figueira *et al.,* 1996). By adding the thiosulfate sulfurtransferase reaction, the *E. coli* WCM should be able to simulate growth in response to cyanide exposure and may be of interest to scientists researching bioleaching for less harmful valuable metal extraction, as

sodium cyanide solution used to leach gold from ore is toxic to wildlife (Eisler & Wiemeyer, 2004). As all the metabolites the reaction needed to be carried out were already present in the metabolites.tsv file, the process of adding a balanced chemical equation to implement the thiosulfate sulfurtransferase reaction into the *E. coli* WCM was simple. The thiosulfate sulfurtransferase reaction is also involved in the transfer of sulfur from thiosulfate to thiophilic acceptors (Ray *et al.,* 2000), but the specific physiological role that this reaction serves is unclear (Libiad *et al.,* 2018). To make more rapid progress in expanding the *E. coli* WCM's representation of metabolism efficiently, we would have to move towards automated inclusion of metabolic reactions from published databases, as opposed to manual curation. Sophie Landon decided to pursue this further, as she had experience with metabolic models. The manual analysis of the *E. coli* WCM's metabolic model I performed also revealed a discrepancy with the EcoCyc data that the model was based on. In the reactions.tsv (https://github.com/JJRightmyer/PrivateWCMEcoli/blob/main/reactions.tsv) file, the xylonate dehydratase reaction (XYLONATE-DEHYDRATASE-RXN) was defined as reversible, contradicting the EcoCyc database (https://ecocyc.org/gene?orgid=ECOLI&id=G7910-MONOMER#tab=RXNS). I was able to inform the model's creators of this inconsistency and they addressed it.

5.3.2 - Implementing transcriptional regulation by Lrp

I worked with Travis Horst to improve transcriptional regulation in the *E. coli* WCM to make it closer to that of an *in vivo* cell. We identified the transcription factor Lrp as a strong candidate for inclusion in the model, as Lrp regulates 38% of *E. coli* genes (Kroner *et al.,* 2019). I was able to download and process the chromatin-immunoprecipitation sequencing (ChIP-seq) and RNA sequencing (RNA-seq) data for 2844 *E. coli* genes published by Kroner *et al.* (2019) into a .tsv file which matched the format of the foldChanges.tsv file (https://github.com/JJRightmyer/PrivateWCMEcoli/blob/main/foldChanges.tsv) in the *E. coli* WCM. This was compatible with the way in which the model calculates transcriptional regulation (Macklin *et al.,* 2020). By calculating the mean of the fold changes in genes across the logarithmic phase (Log), transition point (Trans), and stationary phase (Stat) I was able to provide an average figure for the fold change of each gene regulated by Lrp in an *E. coli* cell grown on minimal media. I then sent the lrp_fold_changes.tsv file I had created to Travis Horst for implementation in the *E. coli* WCM.

5.4 - Discussion

Through collaboration with the creators of the *E. coli* WCM, I was able to contribute to the metabolic model and transcriptional regulation model, implementing the thiosulfate sulfurtransferase reaction and fold change data for 2844 genes regulated by the transcription factor Lrp respectively.

While the addition of a single metabolic reaction to the metabolic model in the *E. coli* WCM represents a small amount of progress, with automated addition of large amounts of data preferred, I increased my familiarity with the model and identified cyanide removal in bioleaching as an enhancement that could increase the potential applications of the model. Manual analysis of the reactions.tsv file also revealed an inconsistency with the EcoCyc database upon which the *E. coli* WCM was based, prompting the model's creators to review data which they had automatically curated. The process of onboarding members of our research group by the Covert Lab was also helpful to them in implementing protocols to make the *E. coli* WCM more accessible to external researchers as they continue to expand their collaborations with external research groups.

My curation of fold change data for 2844 genes by the transcription factor Lrp shall hopefully advance transcriptional regulation within the *E. coli* WCM. Expanding the number of genes in the *E. coli* WCM Lrp regulates from 16 (in foldChanges.tsv) to 2844 should improve the *E. coli* WCM's ability to regulate metabolism and motility in response to changing environmental conditions (Kroner *et al.,* 2019) as the model is developed further. While I was unable to implement the data I curated myself, I gained an understanding of how transcriptional regulation is represented in the *E. coli* WCM and have hopefully provided a useful contribution to the continued development of the model.

Chapter 6 - General Discussion

6.1 - Results Summary

Prior to this Masters project, the *M. genitalium* WCM had been available for use in research since 2012 (Karr *et al.,* 2012), prompting production of algorithms for genome design, resulting in the publication of a new *in silico* minimal genome (Rees-Garbutt *et al.,* 2020c). The difficulty of working with *M. genitalium* in the lab has stifled further progress *in vivo*, leading to researchers investigating other organisms such as *M. mycoides,* creating JCVI-syn3.0 (Hutchison *et al.,* 2016) and JCVI-syn3A (Glass, 2017). *M. pneumoniae* has also been identified as a more suitable organism for *in vivo* research; Piñero-Lambea *et al.* (2020) have developed an oligonucleotide recombineering and Cas-9 mediated counterselection system for genome editing in *M. pneumoniae*, with the intention of making a chassis strain. CRISPR-Cas9 gene editing technology was also well-established in both prokaryotes and eukaryotes (Jinek *et al.,* 2012; Jiang *et al.,* 2013; Hsu *et al.,* 2014), and when combined with λ Red machinery, produced genomic edits in bacteria with relative ease (Reisch & Prather, 2015). Through our group's collaboration with the Covert Lab, we had access to the *E. coli* WCM before its publication in July 2020 (Macklin *et al.,* 2020). Here lay the opportunity to test the *in silico* predictions of a whole-cell model *in vivo*, as *E. coli* is much easier to work with in the laboratory than *M. genitalium* (Hutchison *et al.,* 2016). However, laboratory work was curtailed by the Covid-19 pandemic.

During my Masters project, I was able to: analyse GO terms to provide biological insights into *in silico* minimal gene sets simulated in the *M. genitalium* WCM (Chapter 2), establish the no-SCAR protocol for scarless recombineering of *E. coli* in our laboratory (Chapter 3), compare 1214 single gene knockouts over two generations in the *E. coli* WCM to the essentiality data presented in the Keio collection (Chapter 4), and begin to implement improvements to the current version of the *E. coli* WCM, adding the thiosulfate sulfurtransferase reaction to the metabolic model and transcriptional regulation data by Lrp for 2844 genes (Chapter 5).

Since starting my project, the *E. coli* WCM was published and used to evaluate the literature, which agreed with the majority of curated data and highlighted discrepancies that were investigated, which led to new biological insights (Macklin *et al.,* 2020). The single gene knockout simulations I conducted using the model (Chapter 4) shall help the model's creators understand how well the *E. coli* WCM can predict the effects of gene knockouts and whether the model is fit for purpose for gene essentiality and genome design studies. The addition of the thiosulfate sulfurtransferase reaction and Lrp transcriptional regulation data by me should bring the *E. coli* WCM closer to the representation of an *E. coli* cell *in vivo* and may generate interest in the model by researchers

interested in processes such as bioleaching. Progress has also been reported in JCVI-syn3A, as a reverse genetics approach revealed the seven genes required for normal cell division in this organism (Pelletier *et al.,* 2021). This research took seven years to accomplish, highlighting the incremental nature of progress in synthetic biology. I believe we should adopt *in silico* approaches such as the work described in this thesis using the *M. genitalium* WCM to guide and support *in vivo* research conducted in organisms like JCVI-syn3A, as there is the opportunity to streamline research efforts. This will become more applicable as more WCMs are published (Whole-Cell Modelling; https://www.wholecell.org/models/).

6.2 - Limitations

The progress I made towards implementing the no-SCAR system for genomic editing of *E. coli* was severely limited by the Covid-19 pandemic. I had hoped to test predictions of the *E. coli* WCM *in vivo*, providing an indication of the model's suitability to inform *E. coli* genome engineering. While access to our laboratories has not yet returned to the level we experienced pre-pandemic, I hope this will change over the coming months, permitting implementation of the no-SCAR system to achieve rapid genome editing in *E. coli* (Reisch & Prather, 2015). The methodology I prepared to produce a 1095bp deletion in the *ack* gene reported in this thesis can be readily adapted to permit efficient testing of the *E. coli* WCM's predictions *in vivo*. This work may be continued by Rayane Nunes Lima, a post doc in our collaborator Elibio Rech's laboratory in Brazil, or by future Masters and PhD students joining our research group.

The research I conducted using the *E. coli* WCM was also accompanied by many limitations. It took our research group six months to get simulations running on the *E. coli* WCM due to complications of installing and running the model on the BlueCrystal supercomputer cluster. This has limited the amount of progress we can report from using the model. Supercomputer storage capabilities also limited our research of single gene knockouts using the *E. coli* WCM to two generations, preventing further investigation into some essential genes which produced dividing cells when knocked out. These results can be linked to sub-generational gene expression or large initial metabolite pools sustaining cells despite essential knockouts, but this will have to be investigated. The remote nature of using the *E. coli* WCM has allowed continued research through the Covid-19 pandemic by new PhD and Masters students in our group, enhancing the prospect of further progress being made.

6.3 - Future work

Our group's discovery of MGS Church and Tomita as some of the most minimal *M. genitalium* genomes shown to function so far *in silico* (Rees-Garbutt *et al.,* 2020b) should prompt *in vivo* investigation of these results. As these genomes, along with all other genomes investigated, did not

produce dividing *in silico* cells until genes were reintroduced, it is important that *in vivo* research is conducted (Rees-Garbutt *et al.,* 2020b). This will clarify whether any of the discrepancies are caused by errors in the *M. genitalium* model, aid in understanding the mistakes previous researchers made when designing minimal gene sets, and provide useful information for genome designers in the future. However, *M. genitalium's* unsuitability for laboratory research shall obstruct these efforts (Hutchison *et al.,* 2016). Coupled with the *M. genitalium* WCM's limitations, including long simulation times, using data from other organisms, and ability to only produce single generations of cells, it is difficult to visualise researchers engaging with this work further, as progress will be slow. The publication of a whole-cell biochemical network for *M. genitalium* and the framework used to construct the network could yet rejuvenate interest in this organism (Burke *et al.,* 2020).

Despite only modelling 43% of well-annotated genes in *E. coli*, the *E. coli* WCM represents a vital progression in whole-cell modelling and should be adopted for use in genome-scale research. Macklin *et al.* (2020) used the *E. coli* WCM for comparison of the literature against itself, identifying areas of contradiction which were subsequently investigated. Retrospective analysis of the literature is important to improving our understanding of *E. coli,* but the *E. coli* WCM must provide tangible benefits to reduce the costs associated with *in vivo E. coli* research for its use to be sustained. The accuracy with which the *E. coli* WCM can predict *in vivo* phenotypes must therefore be tested before researchers can begin to optimise design-build-test cycles using the model. An immediate avenue for research may be to simulate 99% knockdowns on all genes in the *E. coli* WCM, as the simulation's phenotype should be very similar to gene knockouts. This will help us understand if the model struggles to compute gene knockouts as opposed to gene knockdowns, despite their similar phenotypes.

Accessibility in the use of the *M. genitalium* and *E. coli* WCMs is key to their continued use among researchers. The data outputs from both models require difficult analysis to extract key information, with machine learning approaches desired but complicated by the nature of the data produced. The onboarding time for new researchers has greatly reduced from the *M. genitalium* to the *E. coli* WCM, due to simpler construction (Macklin *et al.,* 2020), yet the models remain difficult to set up on different supercomputer clusters and remain familiar largely to the teams who develop models, as opposed to the microbiologists, geneticists, and synthetic biologists who could gain from their use.

The complex nature of gene essentiality must also be considered, as the essentiality of some genes is contextual, depending on both the environmental conditions present and their genomic context (Rees-Garbutt *et al.,* 2020c). To accurately reflect the behaviour of genes *in vivo* and provide useful predictions for gene knockout studies, the *E. coli* WCM must be improved to take these factors into

account. The Vivarium Project begins to address the effects of environmental conditions by modelling chemotaxis in *E. coli* and improving the *in silico* cell's ability to interact with heterogenous environments (Agmon & Spangler, 2020). Gene essentiality studies remain relevant when validating cellular models and have featured in recent publications (Macklin *et al.,* 2020; Burke *et al.,* 2020).

As the costs of gene editing, sequencing, and synthesis technologies continue to fall, the fields of genetics and synthetic biology shall continue to enjoy investment and success. CRISPR-Cas9 has been used to produce megabase-sized synthetic genomes (Zhou *et al.,* 2016) and appears to be the premier technology for future genome engineering. While sgRNA design to reduce off-target effects is complex and requires further understanding (Lino *et al.,* 2018), the development of computational tools has simplified this process (Bradford & Perrin, 2019).

The cost of synthesising genomes is also set to decrease, with predictions that synthesising the human genome would cost significantly less than the ~$40,000,000 cost of producing JCVI-syn1.0 (Pennisi, 2010; Schindler *et al.,* 2018). However, the cost of synthesis projects of this scale is still likely to remain obstructive to all but the largest institutes with significant financial support.

Despite costs reducing across synthetic biology, it is possible that the field of whole-cell modelling may stagnate. As new WCMs take years to develop and *in vivo* genome design is restricted to institutions with the greatest funding, the field may become inaccessible to most researchers, with research output too slow and expensive to compete for funding. Multiple reviews have been published addressing these concerns and proposing solutions to keep whole-cell modelling competitive and current. Solutions include: the curation of normalised data for inclusion in WCMs without obtrusive manual effort (Marucci *et al*., 2020), the improvement of experimental data to build models from (Szigeti *et al.,* 2018), and systematic and streamlined construction of WCMs, especially when progressing towards more complex cells such as human (Goldberg *et al.,* 2016). The publication of Datanator by the Karr Lab (Roth *et al.,* 2021) may begin to alleviate the lengthy manual process of WCM data curation and eventually work towards the development of automated WCM construction tools similar to ModelSEED (Seaver *et al.,* 2021) and CarveME (Machado *et al.,* 2018) which already exist for metabolic models. The publication of a *Saccharomyces cerevisiae* WCM (Münzner *et al.,* 2019), alongside WCMs for *M. pneumoniae*, H1 human embryonic stem cell and an archetypal bacterium in development (Whole-Cell Modelling; https://www.wholecell.org/models/), may galvanise further parts of the biological scientific community around these WCMs, securing the future of the field.

6.4 - Conclusions

We are drawing closer to the complete integration of *in silico* and *in vivo* research for custom genome design. The publication of the *E. coli* WCM is important for this goal, providing a WCM for an organism that is both easy to work with in the laboratory and of industrial interest. Before the *E. coli* WCM can be adopted for ubiquitous use by synthetic biologists, its development must be completed, with the remaining 57% of well-annotated genes modelled and missing submodels implemented. I believe a complete *E. coli* WCM will provide us with greater opportunity to investigate the literature as Macklin *et al.* (2020) have done, as it was possible to identify discrepancies within the literature despite the majority of *E. coli* genes being unmodelled. If this progress is achieved and the *E. coli* WCM can accurately predict *in vivo* phenotypes to inform the *in vivo* construction of a genome, we may experience a shift in genome design methodology towards WCM use. This may form the unifying progress whole-cell modelling requires to warrant more widespread use among synthetic biologists.

Tools for *in vivo* genome engineering are suitable and ready for complete design and construction of genomes. The work of the Chin Lab and the JCVI has greatly evidenced this (Fredens *et al.,* 2019; Hutchison *et al.,* 2016). The opportunity to provide genome engineers with accurate data for genome construction is held by the whole-cell modelling community. When WCMs are improved to the point where their use shall streamline genome design in a cost-effective manner, we could experience an explosion of interest in these models by the wider scientific community.

Bibliography

Agmon, E., & Spangler, R. K. (2020). A Multi-Scale Approach to Modeling *E. coli* Chemotaxis. *Entropy, 22,* 1101.

Aite, M., Chevallier, M., Frioux, C., Trottier, C., Got, J., Cortés, M. P., Mendoza, S. N., Carrier, G., Dameron, O., Guillaudeux, N., Latorre, M., Loira, N., Markov, G. V, Maass, A., & Siegel, A. (2018). Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models. *PLOS Computational Biology*, *14*, e1006146.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., & Yeh, L. L. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, *32*, D115–D119.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., & Mori, H. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, *2*, 2006.0008.

Baumgart, M., Unthan, S., Rückert, C., Sivalingam, J., Grünberger, A., Kalinowski, J., Bott, M., Noack, S., & Frunzke, J. (2013). Construction of a prophage-free variant of *Corynebacterium glutamicum* ATCC 13032 for use as a platform strain for basic research and industrial biotechnology. *Applied and Environmental Microbiology*, *79*, 6006–6015.

Benders, G. A., Noskov, V. N., Denisova, E. A., Lartigue, C., Gibson, D. G., Assad-Garcia, N., Chuang, R.-Y., Carrera, W., Moodie, M., Algire, M. A., Phan, Q., Alperovich, N., Vashee, S., Merryman, C., Venter, J. C., Smith, H. O., Glass, J. I., & Hutchison, C. A. (2010). Cloning whole bacterial genomes in yeast. *Nucleic Acids Research*, *38*, 2558–2569.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., & Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science*, *277*, 1453–1462.

Blount, Z. D. (2015). The Natural History of Model Organisms: The unexhausted potential of *E. coli*. *eLife*, *4,* e05826.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., & Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, *326*, 1509–1512.

Bowater, R., & Doherty, A. J. (2006). Making ends meet: Repairing breaks in bacterial DNA by non-homologous end-joining. *PLoS Genetics*, *2*, 93–99.

Bradford, J., & Perrin, D. (2019). Improving CRISPR guide design with consensus approaches. *BMC Genomics*, *20*, 931.

Breuer, M., Earnest, T. M., Merryman, C., Wise, K. S., Sun, L., Lynott, M. R., Hutchison, C. A., Smith, H. O., Lapek, J. D., Gonzalez, D. J., de Crécy-Lagard, V., Haas, D., Hanson, A. D., Labhsetwar, P., Glass, J. I., & Luthey-Schulten, Z. (2019). Essential metabolism for a minimal cell. *eLife*, *8*, e36842.

Burke, P. E. P., Campos, C. B. de L., Costa, L. da F., & Quiles, M. G. (2020). A biochemical network modeling of a whole-cell. *Scientific Reports*, *10*, 13303.

Carr, P. A., & Church, G. M. (2009). Genome engineering. *Nature Biotechnology*, *27*, 1151–1162.

Choi, H. S., Lee, S. Y., Kim, T. Y., & Woo, H. M. (2010). In silico identification of gene amplification targets for improvement of lycopene production. *Applied and Environmental Microbiology*, *76*, 3097–3105.

Datsenko, K. A., & Wanner, B. L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 6640–6645.

de Crécy-Lagard, V., Marck, C., Brochier-Armanet, C., & Grosjean, H. (2007). Comparative RNomics and Modomics in Mollicutes: Prediction of Gene Function and Evolutionary Implications. *IUBMB Life*, *59*, 634-658.

Dzieciol, A. J., & Mann, S. (2012). Designs for life: protocell models in the laboratory. *Chemical Society Reviews*, *41*, 79–85.

Edwards, J. S., & Palsson, B. O. (1999). Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *Journal of Biological Chemistry*, *274*, 17410–17416.

Eisler, R., & Wiemeyer, S. N. (2004). Cyanide hazards to plants and animals from gold mining and related water issues. *Reviews of Environmental Contamination and Toxicology*, *183*, 21–54.

Figueira, M. M., Ciminelli, V. S. T., Andrade, M. C. de, & Linardi, V. R. (1996). Cyanide degradation by an *Escherichia coli* strain. *Canadian Journal of Microbiology*, *42*, 519–523.

Forster, A. C., & Church, G. M. (2006). Towards synthesis of a minimal cell. *Molecular Systems Biology*, *2*, 45.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, R. D., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., … Venter, J. C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, *270*, 397–403.

Fredens, J., Wang, K., de la Torre, D., Funke, L. F. H., Robertson, W. E., Christova, Y., Chia, T., Schmied, W. H., Dunkelmann, D. L., Beránek, V., Uttamapinant, C., Llamazares, A. G., Elliott, T. S., & Chin, J. W. (2019). Total synthesis of *Escherichia coli* with a recoded genome. *Nature*, *569*, 514–518.

Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., Somera, A. L., Kyrpides, N. C., Anderson, I., Gelfand, M. S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M. V., Grechkin, Y., Mseeh, F., Fonstein, M. Y., Overbeek, R., Barabási, A. L., Oltvai, Z. N., … Osterman, A. L. (2003). Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *Journal of Bacteriology*, *185*, 5673–5684.

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. A., Benders, G. A., Montague, M. G., Ma, L., Moodie, M. M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E. A., Young, L., Qi, Z.-Q., Segall-Shapiro, T. H., … Venter, J. C. (2010). Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science*, *329*, 52 – 56.

Gil, R. (2014). The Minimal Gene-Set Machinery. In *Reviews in Cell Biology and Molecular Medicine* (pp. 1–36).

Glass, J. I. (2017). Synthetic bacterium JCVI-Syn3.0 strain 6d, complete genome. *NCBI Nucleotide.* CP016816.2

Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., ... & Venter, J. C. (2006). Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences*, *103*, 425-430.

Glass, J. I., Merryman, C., Wise, K. S., Hutchison, C. A., & Smith, H. O. (2017). Minimal Cells-Real and Imagined. *Cold Spring Harbor Perspectives in Biology*, *9*, a023861.

Goldberg, A. P., Chew, Y. H., & Karr, J. R. (2016). Toward scalable whole-cell modeling of human cells. *SIGSIM-PADS 2016 - Proceedings of the 2016 Annual ACM Conference on Principles of Advanced Discrete Simulation*, 259–262.

Goodall, E. C. A., Robinson, A., Johnston, I. G., Jabbari, S., Turner, K. A., Cunningham, A. F., Lund, P. A., Cole, J. A., & Henderson, I. R. (2018). The Essential Genome of *Escherichia coli*; K-12. *MBio*, *9*, e02096-17.

Göpfrich, K., Platzman, I., & Spatz, J. P. (2018). Mastering Complexity: Towards Bottom-up Construction of Multifunctional Eukaryotic Synthetic Cells. *Trends in Biotechnology*, *36*, 938–951.

Haimovich, A. D., Muir, P., & Isaacs, F. J. (2015). Genomes by design. *Nature Reviews Genetics*, *16*, 501–516.

Hanahan, D. (1985). *DNA cloning: A practical approach* (Vol. 1). McLean, Virginia: IRL Press.

Hanemaaijer, M., Olivier, B. G., Röling, W. F. M., Bruggeman, F. J., & Teusink, B. (2017). Model-based quantification of metabolic interactions from dynamic microbial-community data. *PLOS ONE*, *12*, e0173183.

Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., Fujimitsu, K., Keyamura, K., Ote, T., Yamakawa, T., Yamazaki, Y., Mori, H., Katayama, T., & Kato, J. (2005). Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Molecular Microbiology*, *55*, 137–149.

Hirokawa, Y., Kawano, H., Tanaka-Masuda, K., Nakamura, N., Nakagawa, A., Ito, M., Mori, H., Oshima, T., & Ogasawara, N. (2013). Genetic manipulations restored the growth fitness of reduced-genome *Escherichia coli*. *Journal of Bioscience and Bioengineering*, *116,* 52-58.

Hsu, P. D., Lander, E. S., & Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, *157*, 1262–1278.

Huang, C. H., Hsiang, T., & Trevors, J. T. (2013). Comparative bacterial genomics: defining the minimal core genome. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology*, *103*, 385-398.

Hutchison, C. A., Chuang, R. Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., Pelletier, J. F., Qi, Z. Q., Richter, R. A., Strychalski, E. A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K. S., Smith, H. O., Glass, J. I., … Venter, J. C. (2016). Design and synthesis of a minimal bacterial genome. *Science*, *351*, aad6253.

Hutchison, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O., & Craig Venter, J. (1999). Global Transposon Mutagenesis and a Minimal *Mycoplasma* Genome. *Science*, *286*, 2165 – 2169.

Iwadate, Y., Honda, H., Sato, H., Hashimoto, M., & Kato, J. (2011). Oxidative stress sensitivity of engineered *Escherichia coli* cells with a reduced genome. *FEMS Microbiology Letters*, *322*, 25–33.

Jain, A., Ong, S. P., Chen, W., Medasani, B., Qu, X., Kocher, M., Brafman, M., Petretto, G., Rignanese, G.-M., Hautier, G., Gunter, D., & Persson, K. A. (2015). FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurrency and Computation: Practice and Experience*, *27*, 5037–5059.

Jiang, W., Bikard, D., Cox, D., Zhang, F., & Marraffini, L. A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature Biotechnology*, *31*, 233–239.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*, 816– 821.

Kang, Y., Durfee, T., Glasner, J. D., Qiu, Y., Frisch, D., Winterberg, K. M., & Blattner, F. R. (2004). Systematic mutagenesis of the *Escherichia coli* genome. *Journal of Bacteriology*, *186*, 4921–4930.

Karas, B. J., Jablanovic, J., Sun, L., Ma, L., Goldgof, G. M., Stam, J., Ramon, A., Manary, M. J., Winzeler, E. A., Venter, J. C., Weyman, P. D., Gibson, D. G., Glass, J. I., Hutchison, C. A., Smith, H. O., & Suzuki, Y. (2013). Direct transfer of whole genomes from bacteria to yeast. *Nature Methods*, *10*, 410–412.

Karcagi, I., Draskovits, G., Umenhoffer, K., Fekete, G., Kovács, K., Méhi, O., Balikó, G., Szappanos, B., Györfy, Z., Fehér, T., Bogos, B., Blattner, F. R., Pál, C., Pósfai, G., & Papp, B. (2016). Indispensability of Horizontally Transferred Genes and Its Impact on Bacterial Genome Streamlining. *Molecular Biology and Evolution*, *33*, 1257–1269.

Karlsen, E., Schulz, C., & Almaas, E. (2018). Automated generation of genome-scale metabolic draft reconstructions based on KEGG. *BMC Bioinformatics*, *19*, 467.

Karr, J. R., Sanghvi, J. C., MacKlin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Assad-Garcia, N., Glass, J. I., & Covert, M. W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell*, *150*, 389–401.

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velázquez-Ramírez, D. A., Weaver, D., Collado-Vides, J., Paulsen, I., … Karp, P. D. (2017). The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Research*, *45*, D543–D550.

Khalil A. M. (2020). The genome editing revolution: review. *Journal of Genetic Engineering & Biotechnology*, *18*, 68.

Kim, J., Webb, A. M., Kershner, J. P., Blaskowski, S., & Copley, S. D. (2014). A versatile and highly efficient method for scarless genome editing in *Escherichia coli* and *Salmonella enterica*. *BMC Biotechnology*, *14*, 84.

Kim, Y. G., Cha, J., & Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 1156–1160.

Kolisnychenko, V., Plunkett, G., 3rd, Herring, C. D., Fehér, T., Pósfai, J., Blattner, F. R., & Pósfai, G. (2002). Engineering a reduced *Escherichia coli* genome. *Genome Research*, *12*, 640–647.

Koonin, E. V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, *1*, 127–136.

Kroner, G. M., Wolfe, M. B., & Freddolino, P. L. (2019). *Escherichia coli* Lrp Regulates One-Third of the Genome via Direct, Cooperative, and Indirect Routes. *Journal of Bacteriology*, *201*, e00411-18.

Landon, S., Rees-Garbutt, J., Marucci, L., & Grierson, C. (2019). Genome-driven cell engineering review: *in vivo* and *in silico* metabolic and genome engineering. *Essays in Biochemistry*, *63*, 267–284.

Li, Z., Li, L., Huo, Y., Chen, Z., Zhao, Y., Huang, J., Jian, S., Rong, Z., Wu, D., Gan, J., Hu, X., Li, J., & Xu, X.-W. (2020). Structure-guided protein engineering increases enzymatic activities of the SGNH family esterases. *Biotechnology for Biofuels*, *13*, 107.

Libiad, M., Motl, N., Akey, D. L., Sakamoto, N., Fearon, E. R., Smith, J. L., & Banerjee, R. (2018). Thiosulfate sulfurtransferase-like domain-containing 1 protein interacts with thioredoxin. *The Journal of Biological Chemistry*, *293*, 2675–2686.

Lino, C. A., Harper, J. C., Carney, J. P., & Timlin, J. A. (2018). Delivering CRISPR: a review of the challenges and approaches. *Drug Delivery*, *25*, 1234–1257.

Lu, K., Wu, B., Wang, J., Zhu, W., Nie, H., Qian, J., Huang, W., & Fang, Z. (2018). Blocking amino acid transporter Os AAP 3 improves grain yield by promoting outgrowth buds and increasing tiller number in rice. *Plant Biotechnology Journal, 16*, 1710–1722.

Ma, D., Yang, L., Fleming, R. M. T., Thiele, I., Palsson, B. O., & Saunders, M. A. (2017). Reliable and efficient solution of genome-scale models of Metabolism and macromolecular Expression. *Scientific Reports*, *7*, 40863.

Machado, D., Andrejev, S., Tramontano, M., & Patil, K. R. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*, *46*, 7542–7553.

Macklin, D. N., Ahn-Horst, T. A., Choi, H., Ruggero, N. A., Carrera, J., Mason, J. C., Sun, G., Agmon, E., DeFelice, M. M., Maayan, I., Lane, K., Spangler, R. K., Gillies, T. E., Paull, M. L., Akhter, S., Bray, S. R., Weaver, D. S., Keseler, I. M., Karp, P. D., … Covert, M. W. (2020). Simultaneous cross-

evaluation of heterogeneous *E. coli* datasets via mechanistic simulation. *Science*, *369*, eaav3751.

Martínez-García, E., Nikel, P. I., Aparicio, T., & de Lorenzo, V. (2014). Pseudomonas 2.0: genetic upgrading of *P. putida* KT2440 as an enhanced host for heterologous gene expression. *Microbial Cell Factories*, *13*, 159.

Marucci, L., Barberis, M., Karr, J., Ray, O., Race, P. R., de Souza Andrade, M., Grierson, C., Hoffmann, S. A., Landon, S., Rech, E., Rees-Garbutt, J., Seabrook, R., Shaw, W., & Woods, C. (2020). Computer-Aided Whole-Cell Design: Taking a Holistic Approach by Integrating Synthetic With Systems Biology. *Frontiers in Bioengineering and Biotechnology*, *8*, 1–11.

McCloskey, D., Palsson, B., & Feist, A. M. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Molecular Systems Biology*, *9*, 1–15.

Mendoza, S. N., Olivier, B. G., Molenaar, D., & Teusink, B. (2019). A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biology*, *20*, 158.

Mosberg, J. A., Lajoie, M. J., & Church, G. M. (2010). Lambda Red Recombineering in *Escherichia coli* Occurs Through a Fully Single-Stranded Intermediate. *Genetics*, *186*, 791–799.

Murphy, K. C. (1998). Use of bacteriophage lambda recombination functions to promote gene replacement in *Escherichia coli*. *Journal of Bacteriology*, *180*, 2063–2071.

Münzner, U., Klipp, E., & Krantz, M. (2019). A comprehensive, mechanistically detailed, and executable model of the cell division cycle in *Saccharomyces cerevisiae*. *Nature Communications*, *10*, 1308.

Mushegian, A. R., & Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences*, *93*, 10268-10273.

Oberhardt, M. A., Palsson, B., & Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, *5*, 1–15.

Ochman, H., Gerber, A. S., & Hartl, D. L. (1988). Genetic applications of an inverse polymerase chain reaction. *Genetics*, *120*, 621–623.

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., & Palsson, B. Ø. (2011). A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism--2011. *Molecular Systems Biology*, *7*, 535.

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*, 245–248.

Ostrov, N., Landon, M., Guell, M., Kuznetsov, G., Teramoto, J., Cervantes, N., Zhou, M., Singh, K., Napolitano, M. G., Moosburner, M., Shrock, E., Pruitt, B. W., Conway, N., Goodman, D. B., Gardner, C. L., Tyree, G., Gonzales, A., Wanner, B. L., Norville, J. E., Lajoie, M. J., … Church, G. M. (2016). Design, synthesis, and testing toward a 57-codon genome. *Science*, *353*, 819–822.

Park, M. K., Lee, S. H., Yang, K. S., Jung, S. C., Lee, J. H., & Kim, S. C. (2014). Enhancing recombinant protein production with an *Escherichia coli* host strain lacking insertion sequences. *Applied Microbiology and Biotechnology*, *98*, 6701–6713.

Pelletier, J. F., Sun, L., Wise, K. S., Assad-Garcia, N., Karas, B. J., Deerinck, T. J., Ellisman, M. H., Mershin, A., Gershenfeld, N., Chuang, R.-Y., Glass, J. I., & Strychalski, E. A. (2021). Genetic requirements for cell division in a genomically minimal cell. *Cell*, *184,* 1 – 11.

Pennisi, E. (2010). Synthetic Genome Brings New Life to Bacterium. *Science*, *328*, 958 – 959.

Piñero-Lambea, C., Garcia-Ramallo, E., Martinez, S., Delgado, J., Serrano, L., & Lluch-Senar, M. (2020). *Mycoplasma pneumoniae* Genome Editing Based on Oligo Recombineering and Cas9-Mediated Counterselection. *ACS Synthetic Biology*, *9*, 1693–1704.

Pósfai, G., Plunkett, G., Fehér, T., Frisch, D., Keil, G. M., Umenhoffer, K., Kolisnychenko, V., Stahl, B., Sharma, S. S., de Arruda, M., Burland, V., Harcum, S. W., & Blattner, F. R. (2006). Emergent properties of reduced-genome *Escherichia coli*. *Science, 312*, 1044–1046.

Price, N. D., Reed, J. L., & Palsson, B. (2004). Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nature Reviews Microbiology*, *2*, 886–897.

Quan, J., & Tian, J. (2011). Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. *Nature Protocols*, *6*, 242–251.

Rancati, G., Moffat, J., Typas, A., & Pavelka, N. (2018). Emerging and evolving concepts in gene essentiality. *Nature Reviews Genetics*, *19*, 34–49.

Ray, W. K., Zeng, G., Potters, M. B., Mansuri, A. M., & Larson, T. J. (2000). Characterization of a 12-kilodalton rhodanese encoded by glpE of *Escherichia coli* and its interaction with thioredoxin. *Journal of Bacteriology*, *182*, 2277–2284.

Roberts, T. C., & Morris, K. V. (2013). Not so pseudo anymore: pseudogenes as therapeutic targets. *Pharmacogenomics*, *14*, 2023–2034.

Rees-Garbutt, J., Chalkley, O., Landon, S., Purcell, O., Marucci, L., & Grierson, C. (2020c). Designing minimal genomes using whole-cell models. *Nature Communications*, *11*, 836.

Rees-Garbutt, J., Rightmyer, J., Chalkley, O., Marucci, L., & Grierson, C. (2020b). Testing theoretical minimal genomes using whole-cell models. *BioRxiv*, 2020.03.26.010363.

Rees-Garbutt, J., Rightmyer, J., Karr, J. R., Grierson, C., & Marucci, L. (2020a). Furthering genome design using models and algorithms. *Current Opinion in Systems Biology*, *24*, 120–126.

Reisch, C. R., & Prather, K. L. J. (2015). The no-SCAR (Scarless Cas9 Assisted Recombineering) system for genome editing in *Escherichia coli*. *Scientific Reports*, *5*, 15096.

Reisch, C. R., & Prather, K. (2017). Scarless Cas9 Assisted Recombineering (no-SCAR) in Escherichia coli, an Easy-to-Use System for Genome Editing. *Current Protocols in Molecular Biology*, *117*, 31.8.1–31.8.20.

Reuß, D. R., Altenbuchner, J., Mäder, U., Rath, H., Ischebeck, T., Sappa, P. K., Thürmer, A., Guérin, C., Nicolas, P., Steil, L., Zhu, B., Feussner, I., Klumpp, S., Daniel, R., Commichau, F. M., Völker, U., & Stülke, J. (2017). Large-scale reduction of the *Bacillus subtilis* genome: consequences for the transcriptional network, resource allocation, and metabolism. *Genome Research*, *27*, 289–299.

Roth, Y. D., Lian, Z., Pochiraju, S., Shaikh, B., & Karr, J. R. (2021). Datanator: an integrated database of molecular data for quantitatively modeling cellular behavior. *Nucleic Acids Research*, *49*, D516–D522.

Schindler, D., Dai, J., & Cai, Y. (2018). Synthetic genomics: a new venture to dissect genome fundamentals and engineer new functions. *Current Opinion in Chemical Biology*, *46*, 56–62.

Seaver, S. M. D., Liu, F., Zhang, Q., Jeffryes, J., Faria, J. P., Edirisinghe, J. N., Mundy, M., Chia, N., Noor, E., Beber, M. E., Best, A. A., DeJongh, M., Kimbrel, J. A., D'haeseleer, P., McCorkle, S. R., Bolton, J. R., Pearson, E., Canon, S., Wood-Charlson, E. M., … Henry, C. S. (2021). The

ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Research*, *49*, D575–D588.

Systems Biology UCSD, (2018). Systems Biology UCSD Database. Systems Biology UCSD. http://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms

Szigeti, B., Roth, Y. D., Sekar, J., Goldberg, A. P., Pochiraju, S. C., & Karr, J. R. (2018). A blueprint for human whole-cell modeling. *Current Opinion in Systems Biology*, *7*, 8–15.

Taj, M. K., Samreen, Z., Ling, J. X., Taj, I., Hassan, T. M., & Yunlin, W. (2014). *Escherichia coli* as a model organism. *International Journal of Engineering Research Science & Technology, 3*, 2319-5991.

Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., ... & Hutchison, C. A. (1999). E-CELL: software environment for whole-cell simulation. *Bioinformatics*, *15*, 72-84.

Tsarmpopoulos, I., Gourgues, G., Blanchard, A., Vashee, S., Jores, J., Lartigue, C., & Sirand-Pugnet, P. (2016). In-Yeast Engineering of a Bacterial Genome Using CRISPR/Cas9. *ACS Synthetic Biology*, *5*, 104–109.

Varma, A., & Palsson, B. O. (1994). Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Nature Biotechnology*, *12*, 994–998.

Waltemath, D., & Wolkenhauer, O. (2016). How Modeling Standards, Software, and Initiatives Support Reproducibility in Systems Biology and Systems Medicine. *IEEE Transactions on Bio-Medical Engineering*, *63*, 1999–2006.

Wang, H., Marcišauskas, S., Sánchez, B. J., Domenzain, I., Hermansson, D., Agren, R., Nielsen, J., & Kerkhoven, E. J. (2018). RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLOS Computational Biology*, *14*, e1006541.

Weaver, D. S., Keseler, I. M., Mackie, A., Paulsen, I. T., & Karp, P. D. (2014). A genome-scale metabolic flux model of *Escherichia coli* K–12 derived from the EcoCyc database. *BMC Systems Biology*, *8*, 79.

Xavier, J. C., Patil, K. R., & Rocha, I. (2014). Systems biology perspectives on minimal and simpler cells. *Microbiology and Molecular Biology Reviews*, *78*, 487–509.

Yang, Z.-K., Luo, H., Zhang, Y., Wang, B., & Gao, F. (2019). Pan-genomic analysis provides novel insights into the association of *E. coli* with human host and its minimal genome. *Bioinformatics*, *35*, 1987–1991.

Zhou, J., Wu, R., Xue, X., & Qin, Z. (2016). CasHRA (Cas9-facilitated Homologous Recombination Assembly) method of constructing megabase-sized DNA. *Nucleic Acids Research*, *44*, e124.

Appendices

Appendix 1

Appendix 1. A summary of gene ontology (GO) term reductions and removals from the baseline *M. genitalium* genome to produce the minimal gene set proposed by Forster & Church (2006). Baseline = the unchanged genome of *M. genitalium* in the *M. genitalium* WCM. The number of GO term categories either Unaffected, Reduced, or Removed is listed below the respective columns.

Unaffected	Baseline	ChurchMGS	Reduced	Baseline	ChurchMGS	Removed	Baseline	ChurchMGS
translation	56	56	glycolytic process	11	10	ATP synthesis coupled proton transport	8	0
protein folding	5	5	DNA replication	10	9	cytoadherence to microvasculature mediated by symbiont protein	6	0
cell redox homeostasis	3	3	tRNA processing	5	4	SOS response	6	0
acetyl-CoA biosynthetic process	2	2	DNA topological change	5	4	DNA recombination	4	0
AMP salvage	2	2	transcription DNA-templated	9	7	pathogenesis	4	0
CDP-diacylglycerol biosynthetic process	2	2	phosphoenolpyruvate-dependent sugar phosphotransferase system	4	3	nucleotide-excision repair	3	0
gluconeogenesis	2	2	protein transport	6	4	base-excision repair	2	0
glycine biosynthetic process	2	2	fatty acid biosynthetic process	3	2	lipid metabolic process	2	0
NAD biosynthetic process	2	2	regulation of transcription DNA-templated	3	2	protein secretion	2	0
nucleoside metabolic process	2	2	transport	10	6	response to oxidative stress	2	0
nucleotide biosynthetic process	2	2	carbohydrate metabolic process	5	3	SRP-dependent cotranslational protein targeting to membrane	2	0
one-carbon metabolic process	2	2	cell cycle	4	2	tRNA threonylcarbamoyladenosine modification	2	0

phenylalanyl-tRNA aminoacylation	2	2	glycerol metabolic process	4	2	ATP hydrolysis coupled proton transport	1	0
purine ribonucleoside salvage	2	2	biosynthetic process	2	1	carbohydrate transport	1	0
5-phosphoribose 1-diphosphate biosynthetic process	1	1	DNA-templated transcription initiation	2	1	carboxylic acid metabolic process	1	0
adenine salvage	1	1	protein lipoylation	2	1	cell adhesion	1	0
adhesion of symbiont to host cell	1	1	UMP salvage	2	1	cellular phosphate ion homeostasis	1	0
alanyl-tRNA aminoacylation	1	1	cell division	5	2	cellular protein modification process	1	0
arginyl-tRNA aminoacylation	1	1	ribosome biogenesis	5	2	chromosome segregation	1	0
asparaginyl-tRNA aminoacylation	1	1	pseudouridine synthesis	3	1	chromosome separation	1	0
chromosome condensation	1	1	rRNA processing	3	1	coenzyme A biosynthetic process	1	0
CTP salvage	1	1	DNA repair	8	1	deoxyribonucleotide catabolic process	1	0
cysteinyl-tRNA aminoacylation	1	1	22			deoxyribose phosphate catabolic process	1	0
de novo' CTP biosynthetic process	1	1				DNA restriction-modification system	1	0
deoxyribonucleotide biosynthetic process	1	1				dTMP biosynthetic process	1	0
DNA biosynthetic process	1	1				dTTP biosynthetic process	1	0
DNA replication initiation	1	1				fructose 1,6-bisphosphate metabolic process	1	0
DNA replication synthesis of RNA primer	1	1				heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	1	0
DNA-templated transcription termination	1	1				lipoprotein biosynthetic process	1	0
dTDP biosynthetic process	1	1				mRNA catabolic process	1	0

enterobacterial common antigen biosynthetic process	1	1
FAD biosynthetic process	1	1
FMN biosynthetic process	1	1
fructose 6-phosphate metabolic process	1	1
galactose metabolic process	1	1
glucose catabolic process	1	1
glucose metabolic process	1	1
glutamyl-tRNA aminoacylation	1	1
glycerol catabolic process	1	1
glycerol ether metabolic process	1	1
glycerol-3-phosphate metabolic process	1	1
glycyl-tRNA aminoacylation	1	1
guanosine tetraphosphate biosynthetic process	1	1
histidine biosynthetic process	1	1
histidyl-tRNA aminoacylation	1	1
IMP salvage	1	1
isoleucyl-tRNA aminoacylation	1	1
leucyl-tRNA aminoacylation	1	1
lysyl-tRNA aminoacylation	1	1
membrane lipid biosynthetic process	1	1
metabolic process	1	1
negative regulation of phosphate metabolic process	1	0
phosphate ion transmembrane transport	1	0
protein catabolic process	1	0
protein import	1	0
protein insertion into membrane	1	0
protein repair	1	0
protein targeting	1	0
regulation of carbohydrate metabolic process	1	0
RNA processing	1	0
uracil salvage	1	0
40		

methionine biosynthetic process	1	1
methionyl-tRNA aminoacylation	1	1
mRNA processing	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
organic acid metabolic process	1	1
pentose-phosphate shunt	1	1
pentose-phosphate shunt non-oxidative branch	1	1
phosphate-containing compound metabolic process	1	1
phosphatidylglycerol biosynthetic process	1	1
phospholipid biosynthetic process	1	1
potassium ion transport	1	1
prolyl-tRNA aminoacylation	1	1
protein refolding	1	1
purine nucleotide biosynthetic process	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of DNA replication	1	1
regulation of DNA-templated transcription elongation	1	1
regulation of translation	1	1
regulation of translational fidelity	1	1

removal of superoxide radicals	1	1
response to heat	1	1
riboflavin biosynthetic process	1	1
rRNA catabolic process	1	1
S-adenosylmethionine biosynthetic process	1	1
selenocysteinyl-tRNA(Sec) biosynthetic process	1	1
seryl-tRNA aminoacylation	1	1
sister chromatid cohesion	1	1
tetrahydrofolate interconversion	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
threonyl-tRNA aminoacylation	1	1
transcription antitermination	1	1
tRNA aminoacylation for protein translation	1	1
tRNA modification	1	1
tRNA thio-modification	1	1
tRNA wobble uridine modification	1	1
tryptophanyl-tRNA aminoacylation	1	1
tyrosyl-tRNA aminoacylation	1	1
UDP-glucose metabolic process	1	1
valyl-tRNA aminoacylation	1	1
93		

Appendix 2

Appendix 2. A summary of gene ontology (GO) term reductions and removals from the baseline *M. genitalium* genome to produce the minimal gene set proposed by Gil (2014). Baseline = the unchanged genome of *M. genitalium* in the *M. genitalium* WCM. The number of GO term categories either Unaffected, Reduced, or Removed is listed below the respective columns.

Unaffected	Baseline	GilMGS	Reduced	Baseline	GilMGS	Removed	Baseline	GilMGS
translation	56	56	DNA replication	10	9	cytoadherence to microvasculature mediated by symbiont protein	6	0
glycolytic process	11	11	DNA topological change	5	4	SOS response	6	0
ATP synthesis coupled proton transport	8	8	transcription DNA-templated	9	7	DNA recombination	4	0
protein transport	6	6	phosphoenolpyruvate-dependent sugar phosphotransferase system	4	3	pathogenesis	4	0
protein folding	5	5	fatty acid biosynthetic process	3	2	nucleotide-excision repair	3	0
tRNA processing	5	5	regulation of transcription DNA-templated	3	2	lipid metabolic process	2	0
cell redox homeostasis	3	3	rRNA processing	3	2	response to oxidative stress	2	0
acetyl-CoA biosynthetic process	2	2	transport	10	6	carbohydrate transport	1	0
AMP salvage	2	2	carbohydrate metabolic process	5	3	carboxylic acid metabolic process	1	0
CDP-diacylglycerol biosynthetic process	2	2	ribosome biogenesis	5	3	cell adhesion	1	0
gluconeogenesis	2	2	cell cycle	4	2	cellular phosphate ion homeostasis	1	0
glycine biosynthetic process	2	2	glycerol metabolic process	4	2	cellular protein modification process	1	0
NAD biosynthetic process	2	2	biosynthetic process	2	1	chromosome condensation	1	0
nucleoside metabolic process	2	2	DNA-templated transcription initiation	2	1	chromosome segregation	1	0

nucleotide biosynthetic process	2	2	protein lipoylation	2	1	chromosome separation	1	0
one-carbon metabolic process	2	2	base-excision repair	2	1	deoxyribonucleotide catabolic process	1	0
phenylalanyl-tRNA aminoacylation	2	2	protein secretion	2	1	deoxyribose phosphate catabolic process	1	0
purine ribonucleoside salvage	2	2	tRNA threonylcarbamoyladenosine modification	2	1	DNA restriction-modification system	1	0
SRP-dependent cotranslational protein targeting to membrane	2	2	cell division	5	2	dTMP biosynthetic process	1	0
UMP salvage	2	2	pseudouridine synthesis	3	1	dTTP biosynthetic process	1	0
5-phosphoribose 1-diphosphate biosynthetic process	1	1	DNA repair	8	2	heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	1	0
adenine salvage	1	1	21			lipoprotein biosynthetic process	1	0
adhesion of symbiont to host cell	1	1				mRNA catabolic process	1	0
alanyl-tRNA aminoacylation	1	1				negative regulation of phosphate metabolic process	1	0
arginyl-tRNA aminoacylation	1	1				phosphate ion transmembrane transport	1	0
asparaginyl-tRNA aminoacylation	1	1				protein catabolic process	1	0
ATP hydrolysis coupled proton transport	1	1				protein repair	1	0
coenzyme A biosynthetic process	1	1				regulation of carbohydrate metabolic process	1	0
CTP salvage	1	1				RNA processing	1	0
cysteinyl-tRNA aminoacylation	1	1				sister chromatid cohesion	1	0

de novo' CTP biosynthetic process	1	1
deoxyribonucleotide biosynthetic process	1	1
DNA biosynthetic process	1	1
DNA replication initiation	1	1
DNA replication synthesis of RNA primer	1	1
DNA-templated transcription termination	1	1
dTDP biosynthetic process	1	1
enterobacterial common antigen biosynthetic process	1	1
FAD biosynthetic process	1	1
FMN biosynthetic process	1	1
fructose 1,6-bisphosphate metabolic process	1	1
fructose 6-phosphate metabolic process	1	1
galactose metabolic process	1	1
glucose catabolic process	1	1
glucose metabolic process	1	1
glutamyl-tRNA aminoacylation	1	1
glycerol catabolic process	1	1
glycerol ether metabolic process	1	1
glycerol-3-phosphate metabolic process	1	1
glycyl-tRNA aminoacylation	1	1
guanosine tetraphosphate biosynthetic process	1	1
histidine biosynthetic process	1	1
histidyl-tRNA aminoacylation	1	1
IMP salvage	1	1

30

isoleucyl-tRNA aminoacylation	1	1
leucyl-tRNA aminoacylation	1	1
lysyl-tRNA aminoacylation	1	1
membrane lipid biosynthetic process	1	1
metabolic process	1	1
methionine biosynthetic process	1	1
methionyl-tRNA aminoacylation	1	1
mRNA processing	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
organic acid metabolic process	1	1
pentose-phosphate shunt	1	1
pentose-phosphate shunt non-oxidative branch	1	1
phosphate-containing compound metabolic process	1	1
phosphatidylglycerol biosynthetic process	1	1
phospholipid biosynthetic process	1	1
potassium ion transport	1	1
prolyl-tRNA aminoacylation	1	1
protein import	1	1
protein insertion into membrane	1	1
protein refolding	1	1
protein targeting	1	1
purine nucleotide biosynthetic process	1	1
pyrimidine nucleobase metabolic process	1	1

pyrimidine nucleoside metabolic process	1	1
regulation of DNA replication	1	1
regulation of DNA-templated transcription elongation	1	1
regulation of translation	1	1
regulation of translational fidelity	1	1
removal of superoxide radicals	1	1
response to heat	1	1
riboflavin biosynthetic process	1	1
rRNA catabolic process	1	1
S-adenosylmethionine biosynthetic process	1	1
selenocysteinyl-tRNA(Sec) biosynthetic process	1	1
seryl-tRNA aminoacylation	1	1
tetrahydrofolate interconversion	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
threonyl-tRNA aminoacylation	1	1
transcription antitermination	1	1
tRNA aminoacylation for protein translation	1	1
tRNA modification	1	1
tRNA thio-modification	1	1
tRNA wobble uridine modification	1	1
tryptophanyl-tRNA aminoacylation	1	1
tyrosyl-tRNA aminoacylation	1	1

UDP-glucose metabolic process	1	1
uracil salvage	1	1
valyl-tRNA aminoacylation	1	1
104		

Appendix 3

Appendix 3. A summary of gene ontology (GO) term reductions and removals from the baseline *M. genitalium* genome to produce the minimal gene set proposed by Glass et al. (2006). Baseline = the unchanged genome of *M. genitalium* in the *M. genitalium* WCM. The number of GO term categories either Unaffected, Reduced, or Removed is listed below the respective columns.

Unaffected	Baseline	GlassMGS	Reduced	Baseline	GlassMGS	Removed	Baseline	GlassMGS
translation	56	56	DNA replication	10	9	DNA recombination	4	0
glycolytic process	11	11	transport	10	9	lipid metabolic process	2	0
transcription DNA-templated	9	9	ATP synthesis coupled proton transport	8	7	carboxylic acid metabolic process	1	0
cytoadherence to microvasculature mediated by symbiont protein	6	6	carbohydrate metabolic process	5	4	cellular protein modification process	1	0
protein transport	6	6	ribosome biogenesis	5	4	chromosome condensation	1	0
DNA topological change	5	5	tRNA processing	5	4	chromosome segregation	1	0
protein folding	5	5	phosphoenolpyruvate-dependent sugar phosphotransferase system	4	3	chromosome separation	1	0
pathogenesis	4	4	cell cycle	4	3	coenzyme A biosynthetic process	1	0
cell redox homeostasis	3	3	glycerol metabolic process	4	3	DNA restriction-modification system	1	0
fatty acid biosynthetic process	3	3	pseudouridine synthesis	3	2	dTMP biosynthetic process	1	0
regulation of transcription DNA-templated	3	3	nucleotide-excision repair	3	2	dTTP biosynthetic process	1	0
rRNA processing	3	3	cell division	5	3	phosphate ion transmembrane transport	1	0
acetyl-CoA biosynthetic process	2	2	DNA-templated transcription initiation	2	1	RNA processing	1	0
AMP salvage	2	2	base-excision repair	2	1	sister chromatid cohesion	1	0
biosynthetic process	2	2	SOS response	6	3	14		
CDP-diacylglycerol biosynthetic process	2	2	response to oxidative stress	2	1			
gluconeogenesis	2	2	DNA repair	8	3			

glycine biosynthetic process	2	2	17
NAD biosynthetic process	2	2	
nucleoside metabolic process	2	2	
nucleotide biosynthetic process	2	2	
one-carbon metabolic process	2	2	
phenylalanyl-tRNA aminoacylation	2	2	
protein lipoylation	2	2	
protein secretion	2	2	
purine ribonucleoside salvage	2	2	
SRP-dependent cotranslational protein targeting to membrane	2	2	
tRNA threonylcarbamoyladenosine modification	2	2	
UMP salvage	2	2	
5-phosphoribose 1-diphosphate biosynthetic process	1	1	
adenine salvage	1	1	
adhesion of symbiont to host cell	1	1	
alanyl-tRNA aminoacylation	1	1	
arginyl-tRNA aminoacylation	1	1	
asparaginyl-tRNA aminoacylation	1	1	
ATP hydrolysis coupled proton transport	1	1	
carbohydrate transport	1	1	
cell adhesion	1	1	

cellular phosphate ion homeostasis	1	1
CTP salvage	1	1
cysteinyl-tRNA aminoacylation	1	1
de novo' CTP biosynthetic process	1	1
deoxyribonucleotide biosynthetic process	1	1
deoxyribonucleotide catabolic process	1	1
deoxyribose phosphate catabolic process	1	1
DNA biosynthetic process	1	1
DNA replication initiation	1	1
DNA replication synthesis of RNA primer	1	1
DNA-templated transcription termination	1	1
dTDP biosynthetic process	1	1
enterobacterial common antigen biosynthetic process	1	1
FAD biosynthetic process	1	1
FMN biosynthetic process	1	1
fructose 1,6-bisphosphate metabolic process	1	1
fructose 6-phosphate metabolic process	1	1
galactose metabolic process	1	1
glucose catabolic process	1	1
glucose metabolic process	1	1
glutamyl-tRNA aminoacylation	1	1
glycerol catabolic process	1	1

glycerol ether metabolic process	1	1
glycerol-3-phosphate metabolic process	1	1
glycyl-tRNA aminoacylation	1	1
guanosine tetraphosphate biosynthetic process	1	1
heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	1	1
histidine biosynthetic process	1	1
histidyl-tRNA aminoacylation	1	1
IMP salvage	1	1
isoleucyl-tRNA aminoacylation	1	1
leucyl-tRNA aminoacylation	1	1
lipoprotein biosynthetic process	1	1
lysyl-tRNA aminoacylation	1	1
membrane lipid biosynthetic process	1	1
metabolic process	1	1
methionine biosynthetic process	1	1
methionyl-tRNA aminoacylation	1	1
mRNA catabolic process	1	1
mRNA processing	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
negative regulation of phosphate metabolic process	1	1
organic acid metabolic process	1	1
pentose-phosphate shunt	1	1

pentose-phosphate shunt non-oxidative branch	1	1
phosphate-containing compound metabolic process	1	1
phosphatidylglycerol biosynthetic process	1	1
phospholipid biosynthetic process	1	1
potassium ion transport	1	1
prolyl-tRNA aminoacylation	1	1
protein catabolic process	1	1
protein import	1	1
protein insertion into membrane	1	1
protein refolding	1	1
protein repair	1	1
protein targeting	1	1
purine nucleotide biosynthetic process	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of carbohydrate metabolic process	1	1
regulation of DNA replication	1	1
regulation of DNA-templated transcription elongation	1	1
regulation of translation	1	1
regulation of translational fidelity	1	1
removal of superoxide radicals	1	1
response to heat	1	1

riboflavin biosynthetic process	1	1
rRNA catabolic process	1	1
S-adenosylmethionine biosynthetic process	1	1
selenocysteinyl-tRNA(Sec) biosynthetic process	1	1
seryl-tRNA aminoacylation	1	1
tetrahydrofolate interconversion	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
threonyl-tRNA aminoacylation	1	1
transcription antitermination	1	1
tRNA aminoacylation for protein translation	1	1
tRNA modification	1	1
tRNA thio-modification	1	1
tRNA wobble uridine modification	1	1
tryptophanyl-tRNA aminoacylation	1	1
tyrosyl-tRNA aminoacylation	1	1
UDP-glucose metabolic process	1	1
uracil salvage	1	1
valyl-tRNA aminoacylation	1	1
124		

Appendix 4

Appendix 4. A summary of gene ontology (GO) term reductions and removals from the baseline M. genitalium genome to produce the minimal gene set proposed by Huang et al. (2013). Baseline = the unchanged genome of *M. genitalium* in the *M. genitalium* WCM. The number of GO term categories either Unaffected, Reduced, or Removed is listed below the respective columns.

Unaffected	Baseline	HuangMGS	Reduced	Baseline	HuangMGS	Removed	Baseline	HuangMGS
translation	56	56	DNA replication	10	9	cytoadherence to microvasculature mediated by symbiont protein	6	0
glycolytic process	11	11	protein transport	6	5	SOS response	6	0
cell redox homeostasis	3	3	tRNA processing	5	4	DNA recombination	4	0
acetyl-CoA biosynthetic process	2	2	protein folding	5	4	pathogenesis	4	0
AMP salvage	2	2	DNA topological change	5	4	nucleotide-excision repair	3	0
CDP-diacylglycerol biosynthetic process	2	2	transcription DNA-templated	9	7	lipid metabolic process	2	0
gluconeogenesis	2	2	phosphoenolpyruvate-dependent sugar phosphotransferase system	4	3	base-excision repair	2	0
glycine biosynthetic process	2	2	fatty acid biosynthetic process	3	2	protein secretion	2	0
NAD biosynthetic process	2	2	regulation of transcription DNA-templated	3	2	response to oxidative stress	2	0
nucleoside metabolic process	2	2	transport	10	6	ATP hydrolysis coupled proton transport	1	0
nucleotide biosynthetic process	2	2	carbohydrate metabolic process	5	3	carbohydrate transport	1	0
one-carbon metabolic process	2	2	ribosome biogenesis	5	3	carboxylic acid metabolic process	1	0
phenylalanyl-tRNA aminoacylation	2	2	cell cycle	4	2	cell adhesion	1	0
purine ribonucleoside salvage	2	2	glycerol metabolic process	4	2	cellular phosphate ion homeostasis	1	0

SRP-dependent cotranslational protein targeting to membrane	2	2	DNA-templated transcription initiation	2	1	cellular protein modification process	1	0
5-phosphoribose 1-diphosphate biosynthetic process	1	1	UMP salvage	2	1	chromosome condensation	1	0
adenine salvage	1	1	biosynthetic process	2	1	chromosome segregation	1	0
adhesion of symbiont to host cell	1	1	protein lipoylation	2	1	chromosome separation	1	0
alanyl-tRNA aminoacylation	1	1	tRNA threonylcarbamoyladenosine modification	2	1	coenzyme A biosynthetic process	1	0
arginyl-tRNA aminoacylation	1	1	cell division	5	2	deoxyribonucleotide catabolic process	1	0
asparaginyl-tRNA aminoacylation	1	1	ATP synthesis coupled proton transport	8	3	deoxyribose phosphate catabolic process	1	0
CTP salvage	1	1	pseudouridine synthesis	3	1	DNA restriction-modification system	1	0
cysteinyl-tRNA aminoacylation	1	1	rRNA processing	3	1	dTMP biosynthetic process	1	0
de novo' CTP biosynthetic process	1	1	DNA repair	8	1	dTTP biosynthetic process	1	0
deoxyribonucleotide biosynthetic process	1	1	24			heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	1	0
DNA biosynthetic process	1	1				lipoprotein biosynthetic process	1	0
DNA replication initiation	1	1				mRNA catabolic process	1	0
DNA replication synthesis of RNA primer	1	1				negative regulation of phosphate metabolic process	1	0
DNA-templated transcription termination	1	1				phosphate ion transmembrane transport	1	0
dTDP biosynthetic process	1	1				protein catabolic process	1	0
enterobacterial common antigen biosynthetic process	1	1				protein insertion into membrane	1	0

FAD biosynthetic process	1	1
FMN biosynthetic process	1	1
fructose 1,6-bisphosphate metabolic process	1	1
fructose 6-phosphate metabolic process	1	1
galactose metabolic process	1	1
glucose catabolic process	1	1
glucose metabolic process	1	1
glutamyl-tRNA aminoacylation	1	1
glycerol catabolic process	1	1
glycerol ether metabolic process	1	1
glycerol-3-phosphate metabolic process	1	1
glycyl-tRNA aminoacylation	1	1
guanosine tetraphosphate biosynthetic process	1	1
histidine biosynthetic process	1	1
histidyl-tRNA aminoacylation	1	1
IMP salvage	1	1
isoleucyl-tRNA aminoacylation	1	1
leucyl-tRNA aminoacylation	1	1
lysyl-tRNA aminoacylation	1	1
membrane lipid biosynthetic process	1	1
metabolic process	1	1
methionine biosynthetic process	1	1
methionyl-tRNA aminoacylation	1	1
protein repair	1	0
regulation of carbohydrate metabolic process	1	0
RNA processing	1	0
sister chromatid cohesion	1	0
uracil salvage	1	0
36		

mRNA processing	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
organic acid metabolic process	1	1
pentose-phosphate shunt	1	1
pentose-phosphate shunt non-oxidative branch	1	1
phosphate-containing compound metabolic process	1	1
phosphatidylglycerol biosynthetic process	1	1
phospholipid biosynthetic process	1	1
potassium ion transport	1	1
prolyl-tRNA aminoacylation	1	1
protein import	1	1
protein refolding	1	1
protein targeting	1	1
purine nucleotide biosynthetic process	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of DNA replication	1	1
regulation of DNA-templated transcription elongation	1	1
regulation of translation	1	1
regulation of translational fidelity	1	1
removal of superoxide radicals	1	1
response to heat	1	1

riboflavin biosynthetic process	1	1
rRNA catabolic process	1	1
S-adenosylmethionine biosynthetic process	1	1
selenocysteinyl-tRNA(Sec) biosynthetic process	1	1
seryl-tRNA aminoacylation	1	1
tetrahydrofolate interconversion	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
threonyl-tRNA aminoacylation	1	1
transcription antitermination	1	1
tRNA aminoacylation for protein translation	1	1
tRNA modification	1	1
tRNA thio-modification	1	1
tRNA wobble uridine modification	1	1
tryptophanyl-tRNA aminoacylation	1	1
tyrosyl-tRNA aminoacylation	1	1
UDP-glucose metabolic process	1	1
valyl-tRNA aminoacylation	1	1
95		

Appendix 5

Appendix 5. A summary of gene ontology (GO) term reductions and removals from the baseline M. genitalium genome to produce the minimal gene set proposed by Hutchison et al. (1999). Baseline = the unchanged genome of *M. genitalium* in the *M. genitalium* WCM. The number of GO term categories either Unaffected, Reduced, or Removed is listed below the respective columns.

Unaffected	Baseline	HutchisonMGS	Reduced	Baseline	HutchisonMGS	Removed	Baseline	HutchisonMGS
translation	56	56	carbohydrate metabolic process	5	4	lipid metabolic process	2	0
glycolytic process	11	11	tRNA processing	5	4	cell adhesion	1	0
DNA replication	10	10	phosphoenolpyruvate-dependent sugar phosphotransferase system	4	3	chromosome segregation	1	0
transport	10	10	glycerol metabolic process	4	3	coenzyme A biosynthetic process	1	0
transcription DNA-templated	9	9	DNA repair	8	6	DNA restriction-modification system	1	0
ATP synthesis coupled proton transport	8	8	pathogenesis	4	3	mRNA catabolic process	1	0
protein transport	6	6	cytoadherence to microvasculature mediated by symbiont protein	6	4	phosphate ion transmembrane transport	1	0
DNA topological change	5	5	SOS response	6	4	regulation of carbohydrate metabolic process	1	0
protein folding	5	5	nucleotide-excision repair	3	2	8		
cell redox homeostasis	3	3	cell division	5	3			
fatty acid biosynthetic process	3	3	ribosome biogenesis	5	3			
regulation of transcription DNA-templated	3	3	cell cycle	4	2			
rRNA processing	3	3	DNA-templated transcription initiation	2	1			

SRP-dependent cotranslational protein targeting to membrane	2	2	DNA recombination	4	2
acetyl-CoA biosynthetic process	2	2	pseudouridine synthesis	3	1
AMP salvage	2	2	15		
base-excision repair	2	2			
biosynthetic process	2	2			
CDP-diacylglycerol biosynthetic process	2	2			
gluconeogenesis	2	2			
glycine biosynthetic process	2	2			
NAD biosynthetic process	2	2			
nucleoside metabolic process	2	2			
nucleotide biosynthetic process	2	2			
one-carbon metabolic process	2	2			
phenylalanyl-tRNA aminoacylation	2	2			
protein lipoylation	2	2			
protein secretion	2	2			
purine ribonucleoside salvage	2	2			
response to oxidative stress	2	2			
tRNA threonylcarbamoyladenosine modification	2	2			
UMP salvage	2	2			
5-phosphoribose 1-diphosphate biosynthetic process	1	1			
adenine salvage	1	1			

adhesion of symbiont to host cell	1	1
alanyl-tRNA aminoacylation	1	1
arginyl-tRNA aminoacylation	1	1
asparaginyl-tRNA aminoacylation	1	1
ATP hydrolysis coupled proton transport	1	1
carbohydrate transport	1	1
carboxylic acid metabolic process	1	1
cellular phosphate ion homeostasis	1	1
cellular protein modification process	1	1
chromosome condensation	1	1
chromosome separation	1	1
CTP salvage	1	1
cysteinyl-tRNA aminoacylation	1	1
de novo' CTP biosynthetic process	1	1
deoxyribonucleotide biosynthetic process	1	1
deoxyribonucleotide catabolic process	1	1
deoxyribose phosphate catabolic process	1	1
DNA biosynthetic process	1	1
DNA replication initiation	1	1
DNA replication synthesis of RNA primer	1	1
DNA-templated transcription termination	1	1
dTDP biosynthetic process	1	1

dTMP biosynthetic process	1	1
dTTP biosynthetic process	1	1
enterobacterial common antigen biosynthetic process	1	1
FAD biosynthetic process	1	1
FMN biosynthetic process	1	1
fructose 1,6-bisphosphate metabolic process	1	1
fructose 6-phosphate metabolic process	1	1
galactose metabolic process	1	1
glucose catabolic process	1	1
glucose metabolic process	1	1
glutamyl-tRNA aminoacylation	1	1
glycerol catabolic process	1	1
glycerol ether metabolic process	1	1
glycerol-3-phosphate metabolic process	1	1
glycyl-tRNA aminoacylation	1	1
guanosine tetraphosphate biosynthetic process	1	1
heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	1	1
histidine biosynthetic process	1	1
histidyl-tRNA aminoacylation	1	1
IMP salvage	1	1
isoleucyl-tRNA aminoacylation	1	1
leucyl-tRNA aminoacylation	1	1

lipoprotein biosynthetic process	1	1
lysyl-tRNA aminoacylation	1	1
membrane lipid biosynthetic process	1	1
metabolic process	1	1
methionine biosynthetic process	1	1
methionyl-tRNA aminoacylation	1	1
mRNA processing	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
negative regulation of phosphate metabolic process	1	1
organic acid metabolic process	1	1
pentose-phosphate shunt	1	1
pentose-phosphate shunt non-oxidative branch	1	1
phosphate-containing compound metabolic process	1	1
phosphatidylglycerol biosynthetic process	1	1
phospholipid biosynthetic process	1	1
potassium ion transport	1	1
prolyl-tRNA aminoacylation	1	1
protein catabolic process	1	1
protein import	1	1
protein insertion into membrane	1	1
protein refolding	1	1

protein repair	1	1
protein targeting	1	1
purine nucleotide biosynthetic process	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of DNA replication	1	1
regulation of DNA-templated transcription elongation	1	1
regulation of translation	1	1
regulation of translational fidelity	1	1
removal of superoxide radicals	1	1
response to heat	1	1
riboflavin biosynthetic process	1	1
RNA processing	1	1
rRNA catabolic process	1	1
S-adenosylmethionine biosynthetic process	1	1
selenocysteinyl-tRNA(Sec) biosynthetic process	1	1
seryl-tRNA aminoacylation	1	1
sister chromatid cohesion	1	1
tetrahydrofolate interconversion	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
threonyl-tRNA aminoacylation	1	1

transcription antitermination	1	1
tRNA aminoacylation for protein translation	1	1
tRNA modification	1	1
tRNA thio-modification	1	1
tRNA wobble uridine modification	1	1
tryptophanyl-tRNA aminoacylation	1	1
tyrosyl-tRNA aminoacylation	1	1
UDP-glucose metabolic process	1	1
uracil salvage	1	1
valyl-tRNA aminoacylation	1	1
132		

Appendix 6

Appendix 6. A summary of gene ontology (GO) term reductions and removals from the baseline M. genitalium genome to produce the minimal gene set proposed by Karr et al. (2012). Baseline = the unchanged genome of *M. genitalium* in the *M. genitalium* WCM. The number of GO term categories either Unaffected, Reduced, or Removed is listed below the respective columns.

Unaffected	Baseline	KarrMGS	Reduced	Baseline	KarrMGS	Removed	Baseline	KarrMGS
translation	56	56	DNA replication	10	9	ATP synthesis coupled proton transport	8	0
glycolytic process	11	11	tRNA processing	5	4	SOS response	6	0
cytoadherence to microvasculature mediated by symbiont protein	6	6	transcription DNA-templated	9	7	DNA recombination	4	0
protein transport	6	6	phosphoenolpyruvate-dependent sugar phosphotransferase system	4	3	nucleotide-excision repair	3	0
DNA topological change	5	5	cell cycle	4	3	base-excision repair	2	0
protein folding	5	5	fatty acid biosynthetic process	3	2	lipid metabolic process	2	0
ribosome biogenesis	5	5	regulation of transcription DNA-templated	3	2	response to oxidative stress	2	0
pathogenesis	4	4	carbohydrate metabolic process	5	3	ATP hydrolysis coupled proton transport	1	0
cell redox homeostasis	3	3	cell division	5	3	carbohydrate transport	1	0
rRNA processing	3	3	transport	10	6	carboxylic acid metabolic process	1	0
acetyl-CoA biosynthetic process	2	2	glycerol metabolic process	4	2	cellular phosphate ion homeostasis	1	0
AMP salvage	2	2	DNA-templated transcription initiation	2	1	chromosome condensation	1	0
CDP-diacylglycerol biosynthetic process	2	2	UMP salvage	2	1	chromosome segregation	1	0
gluconeogenesis	2	2	biosynthetic process	2	1	chromosome separation	1	0
glycine biosynthetic process	2	2	protein lipoylation	2	1	coenzyme A biosynthetic process	1	0

NAD biosynthetic process	2	2	protein secretion	2	1	deoxyribonucleotide catabolic process	1	0
nucleoside metabolic process	2	2	pseudouridine synthesis	3	1	deoxyribose phosphate catabolic process	1	0
nucleotide biosynthetic process	2	2	DNA repair	8	1	DNA restriction-modification system	1	0
one-carbon metabolic process	2	2	18			dTMP biosynthetic process	1	0
phenylalanyl-tRNA aminoacylation	2	2				dTTP biosynthetic process	1	0
purine ribonucleoside salvage	2	2				lipoprotein biosynthetic process	1	0
SRP-dependent cotranslational protein targeting to membrane	2	2				mRNA catabolic process	1	0
tRNA threonylcarbamoyladenosine modification	2	2				negative regulation of phosphate metabolic process	1	0
5-phosphoribose 1-diphosphate biosynthetic process	1	1				phosphate ion transmembrane transport	1	0
adenine salvage	1	1				protein repair	1	0
adhesion of symbiont to host cell	1	1				regulation of carbohydrate metabolic process	1	0
alanyl-tRNA aminoacylation	1	1				RNA processing	1	0
arginyl-tRNA aminoacylation	1	1				sister chromatid cohesion	1	0
asparaginyl-tRNA aminoacylation	1	1				uracil salvage	1	0
cell adhesion	1	1				29		
cellular protein modification process	1	1						
CTP salvage	1	1						
cysteinyl-tRNA aminoacylation	1	1						
de novo' CTP biosynthetic process	1	1						

deoxyribonucleotide biosynthetic process	1	1
DNA biosynthetic process	1	1
DNA replication initiation	1	1
DNA replication synthesis of RNA primer	1	1
DNA-templated transcription termination	1	1
dTDP biosynthetic process	1	1
enterobacterial common antigen biosynthetic process	1	1
FAD biosynthetic process	1	1
FMN biosynthetic process	1	1
fructose 1,6-bisphosphate metabolic process	1	1
fructose 6-phosphate metabolic process	1	1
galactose metabolic process	1	1
glucose catabolic process	1	1
glucose metabolic process	1	1
glutamyl-tRNA aminoacylation	1	1
glycerol catabolic process	1	1
glycerol ether metabolic process	1	1
glycerol-3-phosphate metabolic process	1	1
glycyl-tRNA aminoacylation	1	1
guanosine tetraphosphate biosynthetic process	1	1
heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	1	1

histidine biosynthetic process	1	1
histidyl-tRNA aminoacylation	1	1
IMP salvage	1	1
isoleucyl-tRNA aminoacylation	1	1
leucyl-tRNA aminoacylation	1	1
lysyl-tRNA aminoacylation	1	1
membrane lipid biosynthetic process	1	1
metabolic process	1	1
methionine biosynthetic process	1	1
methionyl-tRNA aminoacylation	1	1
mRNA processing	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
organic acid metabolic process	1	1
pentose-phosphate shunt	1	1
pentose-phosphate shunt non-oxidative branch	1	1
phosphate-containing compound metabolic process	1	1
phosphatidylglycerol biosynthetic process	1	1
phospholipid biosynthetic process	1	1
potassium ion transport	1	1
prolyl-tRNA aminoacylation	1	1
protein catabolic process	1	1
protein import	1	1

protein insertion into membrane	1	1
protein refolding	1	1
protein targeting	1	1
purine nucleotide biosynthetic process	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of DNA replication	1	1
regulation of DNA-templated transcription elongation	1	1
regulation of translation	1	1
regulation of translational fidelity	1	1
removal of superoxide radicals	1	1
response to heat	1	1
riboflavin biosynthetic process	1	1
rRNA catabolic process	1	1
S-adenosylmethionine biosynthetic process	1	1
selenocysteinyl-tRNA(Sec) biosynthetic process	1	1
seryl-tRNA aminoacylation	1	1
tetrahydrofolate interconversion	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
threonyl-tRNA aminoacylation	1	1
transcription antitermination	1	1

tRNA aminoacylation for protein translation	1	1
tRNA modification	1	1
tRNA thio-modification	1	1
tRNA wobble uridine modification	1	1
tryptophanyl-tRNA aminoacylation	1	1
tyrosyl-tRNA aminoacylation	1	1
UDP-glucose metabolic process	1	1
valyl-tRNA aminoacylation	1	1
108		

Appendix 7

Appendix 7. A summary of gene ontology (GO) term reductions and removals from the baseline M. genitalium genome to produce the minimal gene set proposed by Mushegian & Koonin (1996). Baseline = the unchanged genome of *M. genitalium* in the *M. genitalium* WCM. The number of GO term categories either Unaffected, Reduced, or Removed is listed below the respective columns.

Unaffected	Baseline	KooninMGS	Reduced	Baseline	KooninMGS	Removed	Baseline	KooninMGS
translation	56	56	protein transport	6	5	cytoadherence to microvasculature mediated by symbiont protein	6	0
glycolytic process	11	11	transcription DNA-templated	9	7	pathogenesis	4	0
DNA replication	10	10	phosphoenolpyruvate-dependent sugar phosphotransferase system	4	3	protein secretion	2	0
ATP synthesis coupled proton transport	8	8	cell cycle	4	3	tRNA threonylcarbamoyladenosine modification	2	0
SOS response	6	6	glycerol metabolic process	4	3	carboxylic acid metabolic process	1	0
DNA topological change	5	5	DNA repair	8	6	cell adhesion	1	0
protein folding	5	5	DNA recombination	4	3	cellular phosphate ion homeostasis	1	0
tRNA processing	5	5	pseudouridine synthesis	3	2	chromosome condensation	1	0
cell redox homeostasis	3	3	fatty acid biosynthetic process	3	2	chromosome segregation	1	0
nucleotide-excision repair	3	3	regulation of transcription DNA-templated	3	2	chromosome separation	1	0
acetyl-CoA biosynthetic process	2	2	rRNA processing	3	2	coenzyme A biosynthetic process	1	0
AMP salvage	2	2	carbohydrate metabolic process	5	3	DNA restriction-modification system	1	0
base-excision repair	2	2	cell division	5	3	heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	1	0
CDP-diacylglycerol biosynthetic process	2	2	DNA-templated transcription initiation	2	1	mRNA catabolic process	1	0

gluconeogenesis	2	2	transport	10	5	negative regulation of phosphate metabolic process	1	0
glycine biosynthetic process	2	2	biosynthetic process	2	1	protein insertion into membrane	1	0
NAD biosynthetic process	2	2	lipid metabolic process	2	1	regulation of carbohydrate metabolic process	1	0
nucleoside metabolic process	2	2	response to oxidative stress	2	1	sister chromatid cohesion	1	0
nucleotide biosynthetic process	2	2	ribosome biogenesis	5	2	18		
one-carbon metabolic process	2	2	19					
phenylalanyl-tRNA aminoacylation	2	2						
protein lipoylation	2	2						
purine ribonucleoside salvage	2	2						
SRP-dependent cotranslational protein targeting to membrane	2	2						
UMP salvage	2	2						
5-phosphoribose 1-diphosphate biosynthetic process	1	1						
adenine salvage	1	1						
adhesion of symbiont to host cell	1	1						
alanyl-tRNA aminoacylation	1	1						
arginyl-tRNA aminoacylation	1	1						
asparaginyl-tRNA aminoacylation	1	1						
ATP hydrolysis coupled proton transport	1	1						
carbohydrate transport	1	1						

cellular protein modification process	1	1
CTP salvage	1	1
cysteinyl-tRNA aminoacylation	1	1
de novo' CTP biosynthetic process	1	1
deoxyribonucleotide biosynthetic process	1	1
deoxyribonucleotide catabolic process	1	1
deoxyribose phosphate catabolic process	1	1
DNA biosynthetic process	1	1
DNA replication initiation	1	1
DNA replication synthesis of RNA primer	1	1
DNA-templated transcription termination	1	1
dTDP biosynthetic process	1	1
dTMP biosynthetic process	1	1
dTTP biosynthetic process	1	1
enterobacterial common antigen biosynthetic process	1	1
FAD biosynthetic process	1	1
FMN biosynthetic process	1	1

fructose 1,6-bisphosphate metabolic process	1	1
fructose 6-phosphate metabolic process	1	1
galactose metabolic process	1	1
glucose catabolic process	1	1
glucose metabolic process	1	1
glutamyl-tRNA aminoacylation	1	1
glycerol catabolic process	1	1
glycerol ether metabolic process	1	1
glycerol-3-phosphate metabolic process	1	1
glycyl-tRNA aminoacylation	1	1
guanosine tetraphosphate biosynthetic process	1	1
histidine biosynthetic process	1	1
histidyl-tRNA aminoacylation	1	1
IMP salvage	1	1
isoleucyl-tRNA aminoacylation	1	1
leucyl-tRNA aminoacylation	1	1
lipoprotein biosynthetic process	1	1
lysyl-tRNA aminoacylation	1	1

membrane lipid biosynthetic process	1	1
metabolic process	1	1
methionine biosynthetic process	1	1
methionyl-tRNA aminoacylation	1	1
mRNA processing	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
organic acid metabolic process	1	1
pentose-phosphate shunt	1	1
pentose-phosphate shunt non-oxidative branch	1	1
phosphate ion transmembrane transport	1	1
phosphate-containing compound metabolic process	1	1
phosphatidylglycerol biosynthetic process	1	1
phospholipid biosynthetic process	1	1
potassium ion transport	1	1
prolyl-tRNA aminoacylation	1	1
protein catabolic process	1	1
protein import	1	1
protein refolding	1	1

protein repair	1	1
protein targeting	1	1
purine nucleotide biosynthetic process	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of DNA replication	1	1
regulation of DNA-templated transcription elongation	1	1
regulation of translation	1	1
regulation of translational fidelity	1	1
removal of superoxide radicals	1	1
response to heat	1	1
riboflavin biosynthetic process	1	1
RNA processing	1	1
rRNA catabolic process	1	1
S-adenosylmethionine biosynthetic process	1	1
selenocysteinyl-tRNA(Sec) biosynthetic process	1	1
seryl-tRNA aminoacylation	1	1
tetrahydrofolate interconversion	1	1
thiamine biosynthetic process	1	1

thiamine diphosphate biosynthetic process	1	1
threonyl-tRNA aminoacylation	1	1
transcription antitermination	1	1
tRNA aminoacylation for protein translation	1	1
tRNA modification	1	1
tRNA thio-modification	1	1
tRNA wobble uridine modification	1	1
tryptophanyl-tRNA aminoacylation	1	1
tyrosyl-tRNA aminoacylation	1	1
UDP-glucose metabolic process	1	1
uracil salvage	1	1
valyl-tRNA aminoacylation	1	1
118		

Appendix 8

Appendix 8. A summary of gene ontology (GO) term reductions and removals from the baseline M. genitalium genome to produce the minimal gene set proposed by Tomita et al. (1999). Baseline = the unchanged genome of *M. genitalium* in the *M. genitalium* WCM. The number of GO term categories either Unaffected, Reduced, or Removed is listed below the respective columns.

Unaffected	Baseline	KooninMGS	Reduced	Baseline	TomitaMGS	Removed	Baseline	TomitaMGS
translation	56	56	DNA replication	10	9	ATP synthesis coupled proton transport	8	0
glycolytic process	11	11	carbohydrate metabolic process	5	4	cytoadherence to microvasculature mediated by symbiont protein	6	0
cell redox homeostasis	3	3	protein folding	5	4	SOS response	6	0
acetyl-CoA biosynthetic process	2	2	DNA topological change	5	4	DNA recombination	4	0
AMP salvage	2	2	tRNA processing	5	4	pathogenesis	4	0
CDP-diacylglycerol biosynthetic process	2	2	transcription DNA-templated	9	7	nucleotide-excision repair	3	0
gluconeogenesis	2	2	phosphoenolpyruvate-dependent sugar phosphotransferase system	4	3	base-excision repair	2	0
glycine biosynthetic process	2	2	glycerol metabolic process	4	3	lipid metabolic process	2	0
NAD biosynthetic process	2	2	fatty acid biosynthetic process	3	2	protein secretion	2	0
nucleoside metabolic process	2	2	regulation of transcription DNA-templated	3	2	response to oxidative stress	2	0
nucleotide biosynthetic process	2	2	protein transport	6	4	SRP-dependent cotranslational protein targeting to membrane	2	0
one-carbon metabolic process	2	2	transport	10	6	tRNA threonylcarbamoyladenosine modification	2	0
phenylalanyl-tRNA aminoacylation	2	2	cell cycle	4	2	ATP hydrolysis coupled proton transport	1	0
purine ribonucleoside salvage	2	2	DNA-templated transcription initiation	2	1	carbohydrate transport	1	0

5-phosphoribose 1-diphosphate biosynthetic process	1	1	biosynthetic process	2	1	cell adhesion	1	0
adenine salvage	1	1	UMP salvage	2	1	cellular phosphate ion homeostasis	1	0
adhesion of symbiont to host cell	1	1	protein lipoylation	2	1	cellular protein modification process	1	0
alanyl-tRNA aminoacylation	1	1	cell division	5	2	chromosome segregation	1	0
arginyl-tRNA aminoacylation	1	1	ribosome biogenesis	5	2	chromosome separation	1	0
asparaginyl-tRNA aminoacylation	1	1	pseudouridine synthesis	3	1	coenzyme A biosynthetic process	1	0
carboxylic acid metabolic process	1	1	rRNA processing	3	1	deoxyribonucleotide catabolic process	1	0
chromosome condensation	1	1	DNA repair	8	1	deoxyribose phosphate catabolic process	1	0
CTP salvage	1	1	22			DNA restriction-modification system	1	0
cysteinyl-tRNA aminoacylation	1	1				dTMP biosynthetic process	1	0
de novo' CTP biosynthetic process	1	1				dTTP biosynthetic process	1	0
deoxyribonucleotide biosynthetic process	1	1				heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	1	0
DNA biosynthetic process	1	1				lipoprotein biosynthetic process	1	0
DNA replication initiation	1	1				mRNA catabolic process	1	0
DNA replication synthesis of RNA primer	1	1				negative regulation of phosphate metabolic process	1	0
DNA-templated transcription termination	1	1				phosphate ion transmembrane transport	1	0
dTDP biosynthetic process	1	1				protein catabolic process	1	0

enterobacterial common antigen biosynthetic process	1	1
FAD biosynthetic process	1	1
FMN biosynthetic process	1	1
fructose 1,6-bisphosphate metabolic process	1	1
fructose 6-phosphate metabolic process	1	1
galactose metabolic process	1	1
glucose catabolic process	1	1
glucose metabolic process	1	1
glutamyl-tRNA aminoacylation	1	1
glycerol catabolic process	1	1
glycerol ether metabolic process	1	1
glycerol-3-phosphate metabolic process	1	1
glycyl-tRNA aminoacylation	1	1
guanosine tetraphosphate biosynthetic process	1	1
histidine biosynthetic process	1	1
histidyl-tRNA aminoacylation	1	1
IMP salvage	1	1
isoleucyl-tRNA aminoacylation	1	1
leucyl-tRNA aminoacylation	1	1
lysyl-tRNA aminoacylation	1	1
protein import	1	0
protein insertion into membrane	1	0
protein refolding	1	0
protein repair	1	0
protein targeting	1	0
regulation of carbohydrate metabolic process	1	0
RNA processing	1	0
uracil salvage	1	0
39		

membrane lipid biosynthetic process	1	1
metabolic process	1	1
methionine biosynthetic process	1	1
methionyl-tRNA aminoacylation	1	1
mRNA processing	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
organic acid metabolic process	1	1
pentose-phosphate shunt	1	1
pentose-phosphate shunt non-oxidative branch	1	1
phosphate-containing compound metabolic process	1	1
phosphatidylglycerol biosynthetic process	1	1
phospholipid biosynthetic process	1	1
potassium ion transport	1	1
prolyl-tRNA aminoacylation	1	1
purine nucleotide biosynthetic process	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of DNA replication	1	1
regulation of DNA-templated transcription elongation	1	1

regulation of translation	1	1
regulation of translational fidelity	1	1
removal of superoxide radicals	1	1
response to heat	1	1
riboflavin biosynthetic process	1	1
rRNA catabolic process	1	1
S-adenosylmethionine biosynthetic process	1	1
selenocysteinyl-tRNA(Sec) biosynthetic process	1	1
seryl-tRNA aminoacylation	1	1
sister chromatid cohesion	1	1
tetrahydrofolate interconversion	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
threonyl-tRNA aminoacylation	1	1
transcription antitermination	1	1
tRNA aminoacylation for protein translation	1	1
tRNA modification	1	1
tRNA thio-modification	1	1
tRNA wobble uridine modification	1	1
tryptophanyl-tRNA aminoacylation	1	1

tyrosyl-tRNA aminoacylation	1	1
UDP-glucose metabolic process	1	1
valyl-tRNA aminoacylation	1	1
94		

Appendix 9

Appendix 9. A summary of gene ontology (GO) terms reintroduced to the *M. genitalium* minimal gene set proposed by Forster & Church (2006) to produce a dividing *in silico* cell.

GOTerms	Baseline	Church Reintroduction
translation	56	4
glycolytic process	11	11
DNA replication	10	8
transport	10	6
transcription DNA-templated	9	7
DNA repair	8	1
protein transport	6	4
carbohydrate metabolic process	5	3
cell division	5	2
DNA topological change	5	4
protein folding	5	4
ribosome biogenesis	5	1
tRNA processing	5	3
cell cycle	4	2
glycerol metabolic process	4	2
phosphoenolpyruvate-dependent sugar phosphotransferase system	4	3
cell redox homeostasis	3	3
pseudouridine synthesis	3	1
regulation of transcription DNA-templated	3	2
rRNA processing	3	1
acetyl-CoA biosynthetic process	2	2
AMP salvage	2	1
biosynthetic process	2	1
CDP-diacylglycerol biosynthetic process	2	2
fatty acid biosynthetic process	2	2
gluconeogenesis	2	2
glycine biosynthetic process	2	2
NAD biosynthetic process	2	2
nucleoside metabolic process	2	2
nucleotide biosynthetic process	2	2
one-carbon metabolic process	2	2
protein lipoylation	2	1
purine ribonucleoside salvage	2	2
UMP salvage	2	1
5-phosphoribose 1-diphosphate biosynthetic process	1	1
adenine salvage	1	1
adhesion of symbiont to host cell	1	1
CTP salvage	1	1
de novo' CTP biosynthetic process	1	1
deoxyribonucleotide biosynthetic process	1	1
DNA biosynthetic process	1	1

DNA replication initiation	1	1
DNA replication synthesis of RNA primer	1	1
DNA-templated transcription initiation	1	1
DNA-templated transcription termination	1	1
dTDP biosynthetic process	1	1
enterobacterial common antigen biosynthetic process	1	1
FAD biosynthetic process	1	1
FMN biosynthetic process	1	1
fructose 1,6-bisphosphate metabolic process	1	1
fructose 6-phosphate metabolic process	1	1
galactose metabolic process	1	1
glucose catabolic process	1	1
glucose metabolic process	1	1
glycerol catabolic process	1	1
glycerol ether metabolic process	1	1
glycerol-3-phosphate metabolic process	1	1
guanosine tetraphosphate biosynthetic process	1	1
histidine biosynthetic process	1	1
IMP salvage	1	1
membrane lipid biosynthetic process	1	1
metabolic process	1	1
methionine biosynthetic process	1	1
mRNA processing	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
organic acid metabolic process	1	1
pentose-phosphate shunt	1	1
pentose-phosphate shunt non-oxidative branch	1	1
phosphatidylglycerol biosynthetic process	1	1
phospholipid biosynthetic process	1	1
potassium ion transport	1	1
purine nucleotide biosynthetic process	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of DNA replication	1	1
regulation of DNA-templated transcription elongation	1	1
regulation of translational fidelity	1	1
removal of superoxide radicals	1	1
response to heat	1	1
riboflavin biosynthetic process	1	1
rRNA catabolic process	1	1
S-adenosylmethionine biosynthetic process	1	1
tetrahydrofolate interconversion	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
transcription antitermination	1	1

tRNA modification	1	1
tRNA thio-modification	1	1
tRNA wobble uridine modification	1	1
UDP-glucose metabolic process	1	1

Appendix 10

Appendix 10. A summary of gene ontology (GO) terms reintroduced to the *M. genitalium* minimal gene set proposed by Gil (2014) to produce a dividing *in silico* cell.

GOTerms	Baseline	Gil Reintroduction
translation	56	6
glycolytic process	11	4
DNA replication	10	1
transport	10	6
transcription DNA-templated	9	2
protein transport	6	4
carbohydrate metabolic process	5	1
cell division	5	1
DNA topological change	5	2
protein folding	5	1
tRNA processing	5	1
cell cycle	4	1
glycerol metabolic process	4	2
cell redox homeostasis	3	2
fatty acid biosynthetic process	3	2
pseudouridine synthesis	3	1
regulation of transcription DNA-templated	3	1
acetyl-CoA biosynthetic process	2	2
AMP salvage	2	1
biosynthetic process	2	1
NAD biosynthetic process	2	2
nucleoside metabolic process	2	1
protein lipoylation	2	1
purine ribonucleoside salvage	2	1
UMP salvage	2	1
adenine salvage	1	1
CTP salvage	1	1
de novo' CTP biosynthetic process	1	1
DNA biosynthetic process	1	1
DNA replication initiation	1	1
enterobacterial common antigen biosynthetic process	1	1
galactose metabolic process	1	1
glycerol catabolic process	1	1
glycerol-3-phosphate metabolic process	1	1
guanosine tetraphosphate biosynthetic process	1	1
membrane lipid biosynthetic process	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
organic acid metabolic process	1	1
phosphatidylglycerol biosynthetic process	1	1
phospholipid biosynthetic process	1	1

potassium ion transport	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of DNA replication	1	1
regulation of translational fidelity	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
tRNA thio-modification	1	1
UDP-glucose metabolic process	1	1

Appendix 11

Appendix 11. A summary of gene ontology (GO) terms reintroduced to the *M. genitalium* minimal gene set proposed by Glass et al. (2006) to produce a dividing *in silico* cell.

GOTerms	Baseline	Glass Reintroduction
glycolytic process	11	1
DNA replication	10	1
transport	10	2
carbohydrate metabolic process	5	1
cell division	5	1
protein folding	5	1
cell cycle	4	1
cell redox homeostasis	3	1
rRNA processing	3	1
CDP-diacylglycerol biosynthetic process	2	1
metabolic process	1	1
mRNA processing	1	1
phosphatidylglycerol biosynthetic process	1	1
rRNA catabolic process	1	1

Appendix 12

Appendix 12. A summary of gene ontology (GO) terms reintroduced to the *M. genitalium* minimal gene set proposed by Huang et al. (2013) to produce a dividing *in silico* cell.

GOTerms	Baseline	Huang Reintroduction
translation	56	25
glycolytic process	11	6
DNA replication	10	4
transcription DNA-templated	9	3
protein transport	6	4
carbohydrate metabolic process	5	2
cell division	5	1
DNA topological change	5	2
protein folding	5	2
ribosome biogenesis	5	1
tRNA processing	5	2
cell cycle	4	1
glycerol metabolic process	4	2
phosphoenolpyruvate-dependent sugar phosphotransferase system	4	2
cell redox homeostasis	3	2
fatty acid biosynthetic process	3	2
pseudouridine synthesis	3	1
regulation of transcription DNA-templated	3	2
acetyl-CoA biosynthetic process	2	2
AMP salvage	2	1
biosynthetic process	2	1
CDP-diacylglycerol biosynthetic process	2	2
glycine biosynthetic process	2	1
NAD biosynthetic process	2	2
nucleoside metabolic process	2	1
nucleotide biosynthetic process	2	1
one-carbon metabolic process	2	1
protein lipoylation	2	1
purine ribonucleoside salvage	2	2
UMP salvage	2	1
adenine salvage	1	1
CTP salvage	1	1
de novo' CTP biosynthetic process	1	1
deoxyribonucleotide biosynthetic process	1	1
DNA biosynthetic process	1	1
enterobacterial common antigen biosynthetic process	1	1
fructose 1,6-bisphosphate metabolic process	1	1
fructose 6-phosphate metabolic process	1	1
galactose metabolic process	1	1
glucose catabolic process	1	1
glycerol catabolic process	1	1

glycerol-3-phosphate metabolic process	1	1
glycyl-tRNA aminoacylation	1	1
guanosine tetraphosphate biosynthetic process	1	1
histidine biosynthetic process	1	1
IMP salvage	1	1
membrane lipid biosynthetic process	1	1
methionine biosynthetic process	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
organic acid metabolic process	1	1
pentose-phosphate shunt non-oxidative branch	1	1
phosphate-containing compound metabolic process	1	1
phosphatidylglycerol biosynthetic process	1	1
phospholipid biosynthetic process	1	1
purine nucleotide biosynthetic process	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of translational fidelity	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
threonyl-tRNA aminoacylation	1	1
tRNA thio-modification	1	1
tRNA wobble uridine modification	1	1
UDP-glucose metabolic process	1	1

Appendix 13

Appendix 13. A summary of gene ontology (GO) terms reintroduced to the *M. genitalium* minimal gene set proposed by Hutchison et al. (1999) to produce a dividing *in silico* cell.

GOTerms	Baseline	Hutchison Reintroduction
translation	56	1
transport	10	1
tRNA processing	5	2
pseudouridine synthesis	3	1
glycine biosynthetic process	2	1
isoleucyl-tRNA aminoacylation	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
tetrahydrofolate interconversion	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
tRNA thio-modification	1	1

Appendix 14

Appendix 14. A summary of gene ontology (GO) terms reintroduced to the *M. genitalium* minimal gene set proposed by Karr et al. (2012) to produce a dividing *in silico* cell.

GOTerms	Baseline	Karr Reintroduction
transport	10	2

Appendix 15

Appendix 15. A summary of gene ontology (GO) terms reintroduced to the *M. genitalium* minimal gene set proposed by Mushegian & Koonin (1996) to produce a dividing *in silico* cell.

GOTerms	Baseline	Koonin Reintroduction
translation	56	5
DNA replication	10	3
transport	10	1
transcription DNA-templated	9	2
carbohydrate metabolic process	5	1
tRNA processing	5	1
glycerol metabolic process	4	1
phosphoenolpyruvate-dependent sugar phosphotransferase system	4	3
fatty acid biosynthetic process	3	1
regulation of transcription DNA-templated	3	2
NAD biosynthetic process	2	1
DNA biosynthetic process	1	1
enterobacterial common antigen biosynthetic process	1	1
membrane lipid biosynthetic process	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
pentose-phosphate shunt non-oxidative branch	1	1
phospholipid biosynthetic process	1	1
potassium ion transport	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of translational fidelity	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
tRNA thio-modification	1	1

Appendix 16

Appendix 16. A summary of gene ontology (GO) terms reintroduced to the *M. genitalium* minimal gene set proposed by Tomita et al. (1999) to produce a dividing *in silico* cell.

GOTerms	Baseline	Tomita Reintroduction
translation	56	7
glycolytic process	11	4
DNA replication	10	8
transport	10	6
transcription DNA-templated	9	3
DNA repair	8	1
protein transport	6	4
carbohydrate metabolic process	5	3
cell division	5	2
DNA topological change	5	4
protein folding	5	4
ribosome biogenesis	5	1
tRNA processing	5	4
cell cycle	4	2
glycerol metabolic process	4	1
cell redox homeostasis	3	3
fatty acid biosynthetic process	3	2
pseudouridine synthesis	3	1
regulation of transcription DNA-templated	3	1
rRNA processing	3	1
acetyl-CoA biosynthetic process	2	2
AMP salvage	2	2
biosynthetic process	2	1
CDP-diacylglycerol biosynthetic process	2	1
glycine biosynthetic process	2	2
NAD biosynthetic process	2	2
nucleoside metabolic process	2	2
nucleotide biosynthetic process	2	2
one-carbon metabolic process	2	2
phenylalanyl-tRNA aminoacylation	2	1
protein lipoylation	2	1
purine ribonucleoside salvage	2	2
UMP salvage	2	1
5-phosphoribose 1-diphosphate biosynthetic process	1	1
adenine salvage	1	1
CTP salvage	1	1
de novo' CTP biosynthetic process	1	1
deoxyribonucleotide biosynthetic process	1	1
DNA biosynthetic process	1	1
DNA replication initiation	1	1
DNA replication synthesis of RNA primer	1	1

DNA-templated transcription termination	1	1
dTDP biosynthetic process	1	1
enterobacterial common antigen biosynthetic process	1	1
FAD biosynthetic process	1	1
FMN biosynthetic process	1	1
galactose metabolic process	1	1
glycerol ether metabolic process	1	1
guanosine tetraphosphate biosynthetic process	1	1
histidine biosynthetic process	1	1
IMP salvage	1	1
membrane lipid biosynthetic process	1	1
metabolic process	1	1
methionine biosynthetic process	1	1
mRNA processing	1	1
NAD metabolic process	1	1
NADP biosynthetic process	1	1
organic acid metabolic process	1	1
pentose-phosphate shunt non-oxidative branch	1	1
phospholipid biosynthetic process	1	1
potassium ion transport	1	1
purine nucleotide biosynthetic process	1	1
pyrimidine nucleobase metabolic process	1	1
pyrimidine nucleoside metabolic process	1	1
regulation of DNA replication	1	1
regulation of DNA-templated transcription elongation	1	1
regulation of translational fidelity	1	1
removal of superoxide radicals	1	1
response to heat	1	1
riboflavin biosynthetic process	1	1
rRNA catabolic process	1	1
S-adenosylmethionine biosynthetic process	1	1
tetrahydrofolate interconversion	1	1
thiamine biosynthetic process	1	1
thiamine diphosphate biosynthetic process	1	1
transcription antitermination	1	1
tRNA modification	1	1
tRNA thio-modification	1	1
tRNA wobble uridine modification	1	1
UDP-glucose metabolic process	1	1

Appendix 17

Appendix 17. Analysis of 1214 single gene knockouts by comparison with the Keio collection. Gene index: Gene index in the *E. coli* WCM, KO number: the corresponding number to knock out a certain gene in the *E. coli* WCM, GeneID: the EcoCyc ID used to identify genes within the *E. coli* WCM, Accession ID: the gene ID used to identify genes within the Keio collection, Gene name: the genes modelled in the *E. coli* WCM (synonyms and identification of pseudogenes are included), Output: a description of whether the single gene knockout simulation produced output (Y = Standard growth output, X = No output, O = Impacted growth output, - = Missing massFractionSummary.png file), Keio essentiality: the essentiality result published by Baba et al. (2006) once separated from values published by Gerdes et al. (2003), Hashimoto et al. (2005), and Kang et al. (2004) (* = weaker confidence in essentiality label by Baba et al. (2006)), Conflicts: whether the output of the single gene knockout conflicts the assessment of essentiality published by Baba et al. (2006) (Y = Conflicts, N = No conflict, X = Unable to assess, Y! = Conflicts but shows impacted growth output, N! = No conflict but shows impacted growth output), Disagreement reason: categorisation of single gene knockouts which conflict the Keio collection (1 = Essential gene produced standard growth output, 2 = Essential gene produced impacted growth output, 3 = Uncertain essentiality gene produced standard growth output, 4 = Non-essential gene produced no output, - = Agrees with the Keio collection).

Gene Index	KO Number	GeneID	AccessionID	Gene name	Output	Keio essentiality	Conflicts	Disagreement reason
15	16	EG10001	ECK4045	alr	Y	N	N	-
18	19	EG10004	ECK3629	dfp	Y	E*	Y	1
19	20	EG10006	ECK4116	dcuB	Y	N	N	-
20	21	EG10007	ECK2301	hisM	Y	N	N	-
24	25	EG10012	ECK0877	cydC	Y	E	Y	1
28	29	EG10016	ECK3114	garL	Y	N	N	-
29	30	EG10017	ECK3176	ispB	Y	E	Y	1
35	36	EG10024	ECK0113	aceE	Y	N	N	-
36	37	EG10025	ECK0114	aceF	Y	N	N	-
38	39	EG10027	ECK2290	ackA	Y	N	N	-
41	42	EG10030	ECK1618	add	Y	N	N	-
42	43	EG10031	ECK1235	adhE	Y	N	N	-
43	44	EG10032	ECK0468	adk	X	E	N	-
45	46	EG10034	ECK2692	alaS	Y	N	N	-
46	47	EG10035	ECK1408	aldA	Y	N	N	-
47	48	EG10036	ECK1295	puuC	Y	N	N	-

49	50	EG10039	ECK1977	amn	Y	N	N	-
53	54	EG10043	ECK4210	cysQ	O	N	N!	-
55	56	EG10045	ECK1765	ansA	Y	N	N	-
56	57	EG10046	ECK2952	ansB	Y	N	N	-
58	59	EG10048	ECK0050	apaH	Y	N	N	-
59	60	EG10049	ECK0971	appA	Y	N*	N	-
61	62	EG10051	ECK0463	apt	Y	N	N	-
62	63	EG10052	ECK0063	araA	Y	u	X	3
63	64	EG10053	ECK0064	araB	Y	u	X	3
64	65	EG10054	ECK0065	araC	Y	N	N	-
65	66	EG10055	ECK0062	araD	Y	u	X	3
66	67	EG10056	ECK2839	araE	Y	N	N	-
67	68	EG10057	ECK1899	araF	Y	N	N	-
68	69	EG10058	ECK1898	araG	Y	N	N	-
69	70	EG10059	ECK1897	araH	Y	N*	N	1
71	72	EG10061	ECK4393	arcA	Y	N	N	1
73	74	EG10063	ECK2814	argA	X	N	Y	4
74	75	EG10064	ECK3950	argB	X	N	Y	4
75	76	EG10065	ECK3949	argC	X	N	Y	4
76	77	EG10066	ECK3347	argD	Y	N	N	-
77	78	EG10067	ECK0274	argF	Y	N	N	-
78	79	EG10068	ECK3161	argG	X	N	Y	4
79	80	EG10069	ECK4247	argI	Y	N	N	-
80	81	EG10070	ECK3226	argR	Y	N	N	-
81	82	EG10071	ECK1877	argS	Y	E	Y	1
82	83	EG10072	ECK2304	argT	Y	N	N	-
83	84	EG10073	ECK0899	aroA	X	N	Y	4
84	85	EG10074	ECK3376	aroB	O	N	N!	-
85	86	EG10075	ECK2323	aroC	O	N	N!	-
86	87	EG10076	ECK1691	aroD	O	N	N!	-

87	88	EG10077	ECK3268	aroE	O	N	N!	-
88	89	EG10078	ECK2598	aroF	Y	N	N	-
89	90	EG10079	ECK0743	aroG	Y	N	N	-
90	91	EG10080	ECK1702	aroH	Y	N	N	-
91	92	EG10081	ECK3377	aroK	Y	N*	N	-
92	93	EG10082	ECK0383	aroL	Y	N	N	-
94	95	EG10084	ECK0111	aroP	Y	N	N	-
98	99	EG10088	ECK3419	asd	O	E	Y!	2
101	102	EG10091	ECK3738	asnA	Y	N	N	-
102	103	EG10092	ECK0662	asnB	Y	N	N	-
104	105	EG10094	ECK0921	asnS	Y	E	Y	1
105	106	EG10095	ECK4133	aspA	Y	N*	N	-
106	107	EG10096	ECK0919	aspC	Y	N	N	-
107	108	EG10097	ECK1867	aspS	Y	E	Y	1
108	109	EG10098	ECK3727	atpA	Y	N	N	-
109	110	EG10099	ECK3731	atpB	Y	N	N	-
110	111	EG10100	ECK3724	atpC	Y	N	N	-
111	112	EG10101	ECK3725	atpD	Y	N	N	-
112	113	EG10102	ECK3730	atpE	Y	N	N	-
113	114	EG10103	ECK3729	atpF	Y	N	N	-
114	115	EG10104	ECK3726	atpG	Y	N	N	-
115	116	EG10105	ECK3728	atpH	Y	N	N	-
117	118	EG10107	ECK3561	avtA	Y	N*	N	-
119	120	EG10109	ECK0309	betA	Y	N	N	-
120	121	EG10110	ECK0310	betB	Y	N	N	-
122	123	EG10112	ECK0312	betT	Y	N	N	-
125	126	EG10115	ECK3715	bglF	Y	N	N	-
127	128	EG10117	ECK0763	bioA	Y	N	N	-
128	129	EG10118	ECK0764	bioB	X	N	Y	4
129	130	EG10119	ECK0766	bioC	Y	N	N	-

130	131	EG10120	ECK0767	bioD	Y	N	N	-
131	132	EG10121	ECK0765	bioF	Y	N	N	-
132	133	EG10122	ECK3399	bioH	Y	N	N	-
133	134	EG10123	ECK3965	birA	Y	E	Y	1
134	135	EG10124	ECK3538	bisC	Y	N	N	-
139	140	EG10129	ECK1708	btuE	Y	N	N	-
141	142	EG10131	ECK4125	cadA	Y	N	N	-
142	143	EG10132	ECK4126	cadB	Y	N	N	-
144	145	EG10134	ECK0033	carA	X	N	Y	4
145	146	EG10135	ECK0034	carB	X	N	Y	4
147	148	EG10137	ECK2136	cdd	Y	N	N	-
148	149	EG10138	ECK3910	cdh	Y	N	N	-
149	150	EG10139	ECK0174	cdsA	Y	E	Y	1
158	159	EG10148	ECK1885	cheR	Y	N	N	-
171	172	EG10161	ECK2043	cpsB	Y	N	N	-
172	173	EG10162	ECK2042	cpsG	Y	N	N	-
175	176	EG10165	ECK2412	crr	Y	N	N	-
180	181	EG10170	ECK3800	cyaA	Y	N	N	-
182	183	EG10173	ECK0721	cydA	Y	E	Y	1
183	184	EG10174	ECK0722	cydB	Y	N	N	-
185	186	EG10176	ECK0336	cynT	Y	N	N	-
187	188	EG10178	ECK0426	cyoA	Y	N	N	-
188	189	EG10179	ECK0425	cyoB	Y	N	N	-
189	190	EG10180	ECK0424	cyoC	Y	N	N	-
190	191	EG10181	ECK0423	cyoD	Y	N	N	-
191	192	EG10182	ECK0422	cyoE	Y	N	N	-
192	193	EG10183	ECK2417	cysA	O	N	N!	-
194	195	EG10185	ECK2745	cysC	O	N	N!	-
195	196	EG10186	ECK2747	cysD	O	N	N!	-
196	197	EG10187	ECK3597	cysE	O	N	N!	-

197	198	EG10188	ECK3356	cysG	Y	N	N	-
198	199	EG10189	ECK2757	cysH	O	N	N!	-
199	200	EG10190	ECK2758	cysI	O	N	N!	-
200	201	EG10191	ECK2759	cysJ	O	N	N!	-
201	202	EG10192	ECK2409	cysK	Y	N	N	-
202	203	EG10193	ECK2416	cysM	Y	N	N	-
203	204	EG10194	ECK2746	cysN	O	N	N!	-
204	205	EG10195	ECK2420	cysP	Y	N	N	-
205	206	EG10196	ECK0519	cysS	Y	E	Y	1
206	207	EG10197	ECK2419	cysU	O	N	N!	-
207	208	EG10198	ECK2418	cysW	O	N*	N!	-
208	209	EG10200	ECK3926	cytR	Y	N	N	-
213	214	EG10205	ECK2474	dapA	O	E	Y!	2
214	215	EG10206	ECK0032	dapB	O	E	Y!	2
215	216	EG10207	ECK0164	dapD	O	E	Y!	2
216	217	EG10208	ECK2467	dapE	O	E	Y!	2
217	218	EG10209	ECK3804	dapF	O	N	N!	-
221	222	EG10213	ECK0376	ddlA	Y	N	N	-
222	223	EG10214	ECK0093	ddlB	Y	N	N	-
225	226	EG10217	ECK2310	accD	Y	E	Y	1
227	228	EG10219	ECK4374	deoA	Y	N	N	-
228	229	EG10220	ECK4375	deoB	Y	N	N	-
229	230	EG10221	ECK4373	deoC	Y	N	N	-
230	231	EG10222	ECK4376	deoD	Y	N	N	-
232	233	EG10224	ECK4034	dgkA	Y	N	N	-
233	234	EG10225	ECK0159	dgt	Y	N	N	-
239	240	EG10231	ECK2126	dld	Y	N	N	-
240	241	EG10232	ECK0885	dmsA	Y	N*	N	-
241	242	EG10233	ECK0886	dmsB	Y	N	N	-
242	243	EG10234	ECK0887	dmsC	X	N	Y	4

243	244	EG10235	ECK3694	dnaA	X	E	N	-
254	255	EG10246	ECK0211	mltD	X	N	Y	4
255	256	EG10247	ECK2561	acpS	Y	E	Y	1
256	257	EG10248	ECK3531	dppA	Y	N	N	-
257	258	EG10249	ECK2362	dsdA	Y	N	N	-
259	260	EG10251	ECK3630	dut	Y	E	Y	1
264	265	EG10256	ECK1851	eda	Y	N	N	-
265	266	EG10257	ECK1852	edd	Y	N	N	-
266	267	EG10258	ECK2773	eno	Y	E	Y	1
267	268	EG10259	ECK0589	entA	Y	N	N	-
269	270	EG10261	ECK0586	entC	Y	N	N	-
270	271	EG10262	ECK0575	entD	Y	E	Y	1
271	272	EG10263	ECK0587	entE	Y	N	N	-
273	274	EG10265	ECK0097	lpxC	Y	E	Y	1
281	282	EG10273	ECK0945	fabA	Y	E	Y	1
282	283	EG10274	ECK2317	fabB	Y	E	Y	1
284	285	EG10276	ECK3243	accC	Y	E	Y	1
285	286	EG10277	ECK1077	fabH	Y	N	N	-
286	287	EG10278	ECK3837	fadA	Y	N*	N	-
287	288	EG10279	ECK3838	fadB	Y	N	N	-
288	289	EG10280	ECK2338	fadL	Y	N	N	-
290	291	EG10282	ECK2921	fbaA	Y	E	Y	1
291	292	EG10283	ECK4227	fbp	Y	N	N	-
293	294	EG10285	ECK4072	fdhF	Y	N	N	-
307	308	EG10299	ECK0577	fes	Y	N	N	-
324	325	EG10316	ECK0178	lpxD	Y	E	Y	1
325	326	EG10317	ECK3248	fis	Y	N	N	-
333	334	EG10325	ECK1330	fnr	Y	N	N	-
334	335	EG10326	ECK0049	folA	Y	E	Y	1
335	336	EG10327	ECK2309	folC	Y	E	Y	1

336	337	EG10328	ECK0522	folD	Y	E	Y	1
338	339	EG10330	ECK4150	frdA	Y	N	N	-
339	340	EG10331	ECK4149	frdB	Y	N	N	-
340	341	EG10332	ECK4148	frdC	Y	N	N	-
341	342	EG10333	ECK4147	frdD	Y	N	N	-
342	343	EG10334	ECK3836	fre	Y	N	N	-
352	353	EG10344	ECK0090	ftsW	Y	E	Y	1
356	357	EG10348	ECK2795	fucA	Y	N	N	-
357	358	EG10349	ECK2797	fucI	Y	N	N	-
358	359	EG10350	ECK2798	fucK	Y	N	N	-
359	360	EG10351	ECK2794	fucO	Y	N	N	-
360	361	EG10352	ECK2796	fucP	Y	N	N	-
362	363	EG10355	ECK2799	fucU	X	N	Y	4
363	364	EG10356	ECK1607	fumA	Y	N	N	-
364	365	EG10357	ECK4115	fumB	Y	N	N	-
365	366	EG10358	ECK1606	fumC	Y	N	N	-
368	369	EG10361	ECK2656	gabT	Y	N	N	-
369	370	EG10362	ECK0748	galE	Y	N*	N	-
370	371	EG10363	ECK0746	galK	Y	N	N	-
373	374	EG10366	ECK0747	galT	Y	N	N	-
374	375	EG10367	ECK1777	gapA	Y	E	Y	1
375	376	EG10368	ECK2923	epd	Y	N	N	-
377	378	EG10370	ECK2511	ispG	Y	E	Y	1
379	380	EG10372	ECK1759	gdhA	Y	N	N	-
386	387	EG10379	ECK3416	glgC	Y	N	N	-
389	390	EG10382	ECK3722	glmS	Y	E	Y	1
390	391	EG10383	ECK3863	glnA	X	N	Y	4
393	394	EG10386	ECK0800	glnH	Y	N	N	-
395	396	EG10388	ECK0799	glnP	Y	N	N	-
396	397	EG10389	ECK0798	glnQ	Y	N	N	-

397	398	EG10390	ECK0668	glnS	Y	E	Y	1
398	399	EG10391	ECK2233	glpA	Y	N	N	-
399	400	EG10392	ECK2234	glpB	Y	N	N	-
400	401	EG10393	ECK2235	glpC	Y	N	N	-
401	402	EG10394	ECK3412	glpD	Y	N	N	-
402	403	EG10395	ECK3411	glpE	Y	N	N	-
403	404	EG10396	ECK3919	glpF	Y	N	N	-
405	406	EG10398	ECK3918	glpK	Y	N	N	-
406	407	EG10399	ECK2231	glpQ	Y	N	N	-
409	410	EG10402	ECK0709	gltA	Y	N	N	-
410	411	EG10403	ECK3202	gltB	Y	N	N	-
411	412	EG10404	ECK3203	gltD	Y	N	N	-
412	413	EG10405	ECK4070	gltP	Y	N	N	-
414	415	EG10407	ECK2394	gltX	-	E	X	N/A
415	416	EG10408	ECK2548	glyA	Y	N	N	-
416	417	EG10409	ECK3548	glyQ	Y	E	Y	1
417	418	EG10410	ECK3547	glyS	Y	N	N	-
418	419	EG10411	ECK2024	gnd	Y	N	N	-
419	420	EG10412	ECK3485	gor	-	N	X	N/A
420	421	EG10413	ECK3771	gpp	Y	N*	N	-
421	422	EG10414	ECK0239	gpt	Y	N	N	-
425	426	EG10418	ECK2683	gshA	Y	N	N	-
426	427	EG10419	ECK2942	gshB	X	N	Y	4
427	428	EG10420	ECK2503	guaA	X	N	Y	4
428	429	EG10421	ECK2504	guaB	Y	N	N	-
429	430	EG10422	ECK0104	guaC	X	N	Y	4
432	433	EG10425	ECK1614	hdhA	Y	N	N	-
434	435	EG10427	ECK1198	hemA	X	E	N	-
435	436	EG10428	ECK0366	hemB	X	E*	N	-
436	437	EG10429	ECK3799	hemC	X	E*	N	-

437	438	EG10430	ECK3798	hemD	X	E	N	-
439	440	EG10432	ECK0153	hemL	Y	E	Y	1
447	448	EG10440	ECK1710	ihfA	Y	N	N	-
448	449	EG10441	ECK0903	ihfB	X	N	Y	4
451	452	EG10444	ECK2019	hisA	O	N	N!	-
452	453	EG10445	ECK2017	hisB	O	N	N!	-
453	454	EG10446	ECK2016	hisC	O	N	N!	-
454	455	EG10447	ECK2015	hisD	O	N	N!	-
455	456	EG10448	ECK2020	hisF	O	N	N!	-
456	457	EG10449	ECK2014	hisG	O	N	N!	-
457	458	EG10450	ECK2018	hisH	O	N	N!	-
458	459	EG10451	ECK2021	hisI	O	N	N!	-
459	460	EG10452	ECK2300	hisP	Y	N	N	-
460	461	EG10453	ECK2510	hisS	X	E	N	-
464	465	EG10457	ECK1232	hns	Y	N	N	-
471	472	EG10464	ECK1039	lpxL	Y	N	N	-
472	473	EG10465	ECK3032	ribB	Y	E	Y	1
482	483	EG10475	ECK2719	hycB	Y	N	N	-
483	484	EG10476	ECK2718	hycC	Y	N	N	-
484	485	EG10477	ECK2717	hycD	Y	N	N	-
485	486	EG10478	ECK2716	hycE	X	N	Y	4
486	487	EG10479	ECK2715	hycF	Y	N	N	-
487	488	EG10480	ECK2714	hycG	Y	N	N	-
496	497	EG10489	ECK1122	icd	X	N	Y	4
497	498	EG10490	ECK2912	argP	Y	N	N	-
499	500	EG10492	ECK0027	ileS	Y	N	N	-
500	501	EG10493	ECK3764	ilvA	Y	N	N	-
501	502	EG10494	ECK3662	ilvB	Y	N	N	-
502	503	EG10495	ECK3766	ilvC	O	N	N!	-
503	504	EG10496	ECK3763	ilvD	O	N*	N!	-

504	505	EG10497	ECK3762	ilvE	O	N*	N!	-
505	506	EG10498	ECK3760	ilvG_2 (pseudo)	X	N*	Y	4
506	507	EG10499	ECK0080	ilvH	Y	N	N	-
507	508	EG10500	ECK0079	ilvI	Y	N*	N	-
508	509	EG10501	ECK3761	ilvM	Y	N	N	-
509	510	EG10502	ECK3661	ilvN	Y	N	N	-
515	516	EG10508	ECK0415	ispA	X	E	N	-
516	517	EG10509	ECK1730	katE	Y	N*	N	-
518	519	EG10511	ECK3934	katG	Y	N	N	-
519	520	EG10512	ECK3607	kbl	Y	N	N	-
520	521	EG10513	ECK0686	kdpA	Y	N	N	-
521	522	EG10514	ECK0685	kdpB	Y	N	N	-
522	523	EG10515	ECK0684	kdpC	Y	N	N	-
525	526	EG10518	ECK1203	kdsA	Y	E	Y	1
526	527	EG10519	ECK0909	kdsB	Y	E	Y	1
527	528	EG10520	ECK3623	waaA/kdtA	Y	E	Y	1
528	529	EG10521	ECK0048	kefC	Y	N	N	-
531	532	EG10524	ECK0339	lacA	Y	N	N	-
533	534	EG10526	ECK0340	lacY	Y	N	N	-
534	535	EG10527	ECK0341	lacZ	Y	u	X	3
538	539	EG10531	ECK0078	leuO	Y	N*	N	-
539	540	EG10532	ECK0635	leuS	Y	E	Y	1
540	541	EG10533	ECK4035	lexA	Y	E	Y	1
543	544	EG10536	ECK3438	livF	Y	N*	N	-
544	545	EG10537	ECK3439	livG	Y	N	N	-
545	546	EG10538	ECK3441	livH	Y	N	N	-
546	547	EG10539	ECK3444	livJ	Y	N*	N	-
547	548	EG10540	ECK3442	livK	Y	N	N	-
548	549	EG10541	ECK3440	livM	Y	N	N	-

550	551	EG10543	ECK0115	lpd	X	N	Y	4
552	553	EG10545	ECK0180	lpxA	Y	E	Y	1
553	554	EG10546	ECK0181	lpxB	Y	E	Y	1
554	555	EG10547	ECK0880	lrp	Y	N	N	-
556	557	EG10549	ECK2836	lysA	O	N	N!	-
557	558	EG10550	ECK4016	lysC	Y	N	N	-
559	560	EG10552	ECK2885	lysS	Y	N	N	-
560	561	EG10553	ECK4123	lysU	Y	N	N	-
567	568	EG10560	ECK3404	malP	Y	N*	N	-
568	569	EG10561	ECK3403	malQ	Y	N	N	-
571	572	EG10564	ECK1617	malY	Y	N	N	-
572	573	EG10565	ECK0397	malZ	Y	N	N	-
573	574	EG10566	ECK1608	manA	Y	N*	N	-
574	575	EG10567	ECK1815	manX	Y	N	N	-
575	576	EG10568	ECK1816	manY	X	N	Y	4
576	577	EG10569	ECK1817	manZ	Y	N*	N	-
579	580	EG10572	ECK2775	mazG	Y	N	N	-
583	584	EG10576	ECK3225	mdh	Y	N	N	-
586	587	EG10579	ECK2258	menD	Y	N	N	-
588	589	EG10581	ECK4005	metA	O	N	N!	-
589	590	EG10582	ECK3931	metB	O	N	N!	-
590	591	EG10583	ECK3000	metC	Y	N	N	-
591	592	EG10584	ECK3823	metE	Y	N	N	-
592	593	EG10585	ECK3933	metF	O	N	N!	-
593	594	EG10586	ECK2107	metG	Y	E	Y	1
594	595	EG10587	ECK4011	metH	Y	N	N	-
595	596	EG10588	ECK3930	metJ	Y	N	N	-
596	597	EG10589	ECK2937	metK	Y	E	Y	1
597	598	EG10590	ECK3932	metL	Y	N	N	-
602	603	EG10595	ECK4167	miaA	Y	N	N	-

611	612	EG10604	ECK0088	mraY	Y	E	Y	1
620	621	EG10613	ECK0905	msbA	Y	E	Y	1
621	622	EG10614	ECK1856	lpxM	Y	N	N	-
623	624	EG10616	ECK3589	mtlD	Y	N	N	-
624	625	EG10617	ECK3149	mtr	Y	N	N	-
626	627	EG10619	ECK0092	murC	Y	E	Y	1
627	628	EG10620	ECK0089	murD	Y	E	Y	1
628	629	EG10621	ECK0086	murE	Y	E	Y	1
629	630	EG10622	ECK0087	murF	Y	E	Y	1
630	631	EG10623	ECK0091	murG	Y	E	Y	1
635	636	EG10628	ECK3916	fpr	Y	N	N	-
636	637	EG10629	ECK0534	emrE	Y	N	N	-
637	638	EG10630	ECK0739	nadA	Y	N	N	-
638	639	EG10631	ECK2572	nadB	Y	N	N	-
639	640	EG10632	ECK0665	nagA	Y	N	N	-
641	642	EG10634	ECK0663	umpH/nagD	Y	N	N	-
644	645	EG10637	ECK3214	nanA	Y	N	N	-
649	650	EG10642	ECK1217	narK	Y	N	N	-
650	651	EG10643	ECK1215	narL	Y	N	N	-
656	657	EG10649	ECK1095	ndh	Y	N	N	-
657	658	EG10650	ECK2514	ndk	X	N	Y	4
659	660	EG10652	ECK0020	nhaA	Y	N	N	-
660	661	EG10653	ECK3353	nirB	Y	N	N	-
661	662	EG10654	ECK3355	nirC	Y	N*	N	-
662	663	EG10655	ECK3354	nirD	Y	N	N	-
667	668	EG10660	ECK2226	nrdA	Y	E	Y	1
668	669	EG10661	ECK2227	nrdB	Y	E	Y	1
670	671	EG10663	ECK1738	nadE	Y	E	Y	1
671	672	EG10664	ECK2959	nupG	Y	N*	N	-
677	678	EG10670	ECK2207	ompC	Y	N	N	-

678	679	EG10671	ECK0920	ompF	Y	N	N	-
681	682	EG10674	ECK1237	oppA	Y	N	N	-
682	683	EG10675	ECK1238	oppB	Y	N	N	-
683	684	EG10676	ECK1239	oppC	Y	N	N	-
684	685	EG10677	ECK1240	oppD	Y	N*	N	-
685	686	EG10678	ECK1241	oppF	Y	N	N	-
689	690	EG10682	ECK3348	pabA	Y	N	N	-
690	691	EG10683	ECK1810	pabB	Y	N	N	-
695	696	EG10688	ECK3390	pck	Y	N	N	-
698	699	EG10691	ECK0053	pdxA	Y	N	N	-
699	700	EG10693	ECK2562	pdxJ	Y	N	N	-
700	701	EG10694	ECK4253	pepA	Y	N	N	-
701	702	EG10695	ECK0238	pepD	Y	N	N	-
702	703	EG10696	ECK0923	pepN	Y	N	N	-
705	706	EG10699	ECK3908	pfkA	Y	N	N	-
706	707	EG10700	ECK1721	pfkB	Y	N*	N	-
708	709	EG10702	ECK4017	pgi	Y	N	N	-
709	710	EG10703	ECK2922	pgk	Y	E	Y	1
710	711	EG10704	ECK0412	pgpA	Y	N	N	-
711	712	EG10705	ECK1273	pgpB	Y	N	N	-
712	713	EG10706	ECK1911	pgsA	Y	E	Y	1
713	714	EG10707	ECK2596	pheA	O	N	N!	-
714	715	EG10708	ECK0568	pheP	Y	N	N	-
715	716	EG10709	ECK1712	pheS	Y	E	Y	1
716	717	EG10710	ECK1711	pheT	Y	E	Y	1
719	720	EG10713	ECK4099	phnC	Y	N	N	-
720	721	EG10714	ECK4098	phnD	Y	N	N	-
729	730	EG10723	ECK4087	phnN	Y	N	N	-
730	731	EG10724	ECK4086	phnO	Y	N	N	-
732	733	EG10727	ECK0378	phoA	Y	N*	N	-

734	735	EG10729	ECK0242	phoE	Y	N	N	-
739	740	EG10734	ECK3721	pstS	Y	N	N	-
744	745	EG10739	ECK3819	pldB	Y	N*	N	-
745	746	EG10740	ECK4033	plsB	Y	E*	Y	1
747	748	EG10742	ECK0922	pncB	Y	N	N	-
748	749	EG10743	ECK3152	pnp	Y	N*	N	-
749	750	EG10744	ECK1598	pntA	Y	N	N	-
750	751	EG10745	ECK1597	pntB	Y	N	N	-
751	752	EG10746	ECK3855	polA	Y	N	N	-
754	755	EG10749	ECK1112	potA	Y	N	N	-
755	756	EG10750	ECK1111	potB	Y	N	N	-
756	757	EG10751	ECK1110	potC	Y	N	N	-
757	758	EG10752	ECK1109	potD	Y	N	N	-
758	759	EG10753	ECK0680	potE	Y	N	N	-
759	760	EG10754	ECK0862	poxB	Y	N	N	-
760	761	EG10755	ECK4222	ppa	X	E	N	-
761	762	EG10756	ECK3947	ppc	Y	N	N	-
762	763	EG10757	ECK3351	ppiA	Y	N	N	-
763	764	EG10758	ECK0518	ppiB	Y	N	N	-
764	765	EG10759	ECK1700	ppsA/pps	Y	N	N	-
772	773	EG10767	ECK0244	proA	Y	N	N	-
773	774	EG10768	ECK0243	proB	Y	N	N	-
774	775	EG10769	ECK0381	proC	O	N	N!	-
775	776	EG10770	ECK0194	proS	Y	E	Y	1
776	777	EG10771	ECK2671	proV	Y	N	N	-
777	778	EG10772	ECK2672	proW	Y	N	N	-
778	779	EG10773	ECK2673	proX	Y	N	N	-
779	780	EG10774	ECK1195	prs/prsA	Y	E	Y	1
786	787	EG10781	ECK2583	pssA	Y	E*	Y	1
787	788	EG10782	ECK3719	pstA	X	N	Y	4

788	789	EG10783	ECK3718	pstB	Y	N	N	-
789	790	EG10784	ECK3720	pstC	Y	N	N	-
792	793	EG10787	ECK1087	ptsG	Y	N	N	-
795	796	EG10790	ECK4173	purA	X	N	Y	4
796	797	EG10791	ECK2472	purC	X	N	Y	4
797	798	EG10792	ECK3997	purD	X	N	Y	4
798	799	EG10793	ECK0516	purE	X	N	Y	4
799	800	EG10794	ECK2306	purF	X	N	Y	4
800	801	EG10795	ECK3998	purH	X	N	Y	4
801	802	EG10796	ECK0515	purK	X	N	Y	4
802	803	EG10797	ECK2555	purL	X	N	Y	4
803	804	EG10798	ECK2495	purM	X	N	Y	4
804	805	EG10799	ECK2496	purN	Y	N	N	-
806	807	EG10801	ECK1005	putA	Y	N	N	-
807	808	EG10802	ECK1006	putP	Y	N	N	-
808	809	EG10803	ECK1855	pykA	Y	N	N	-
809	810	EG10804	ECK1672	pykF	Y	N	N	-
810	811	EG10805	ECK4240	pyrB	Y	N	N	-
811	812	EG10806	ECK1047	pyrC	Y	N	N	-
812	813	EG10807	ECK0936	pyrD	Y	N	N	-
813	814	EG10808	ECK3632	pyrE	Y	E	Y	1
814	815	EG10809	ECK1276	pyrF	Y	N	N	-
815	816	EG10810	ECK2774	pyrG	X	E	N	-
816	817	EG10811	ECK4239	pyrI	Y	N	N	-
822	823	EG10817	ECK3742	rbsD	Y	N	N	-
826	827	EG10821	ECK2210	rcsB	Y	N	N	-
840	841	EG10835	ECK2778	relA	Y	N	N	-
843	844	EG10838	ECK3609	rfaD	Y	N	N	-
849	850	EG10844	ECK3772	rhlB	Y	N	N	-
860	861	EG10855	ECK0634	lptE/rlpB	Y	E	Y	1

861	862	EG10856	ECK0604	rna	Y	N	N	-
862	863	EG10857	ECK2565	rnc	Y	E	Y	1
863	864	EG10858	ECK1802	rnd	Y	N	N	-
864	865	EG10859	ECK1069	rne	Y	E	Y	1
865	866	EG10860	ECK0214	rnhA	Y	N	N	-
866	867	EG10861	ECK0182	rnhB	Y	N	N	-
867	868	EG10862	ECK3696	rnpA	Y	E	Y	1
868	869	EG10863	ECK3633	rph	Y	N	N	-
869	870	EG10864	ECK3975	rplA	X	N	Y	4
870	871	EG10865	ECK3304	rplB	X	E	N	-
871	872	EG10866	ECK3307	rplC	X	E	N	-
872	873	EG10867	ECK3306	rplD	X	E	N	-
873	874	EG10868	ECK3295	rplE	X	E	N	-
874	875	EG10869	ECK3292	rplF	X	E	N	-
875	876	EG10870	ECK4199	rplI	X	N	Y	4
876	877	EG10871	ECK3976	rplJ	X	E	N	-
877	878	EG10872	ECK3974	rplK	X	N	Y	4
878	879	EG10873	ECK3977	rplL	X	E	N	-
879	880	EG10874	ECK3220	rplM	X	E	N	-
880	881	EG10875	ECK3297	rplN	X	E	N	-
881	882	EG10876	ECK3288	rplO	X	E	N	-
882	883	EG10877	ECK3300	rplP	X	E	N	-
883	884	EG10878	ECK3281	rplQ	X	E	N	-
884	885	EG10879	ECK3291	rplR	X	E	N	-
885	886	EG10880	ECK2603	rplS	X	E	N	-
886	887	EG10881	ECK1714	rplT	X	E	N	-
887	888	EG10882	ECK3302	rplV	X	E	N	-
888	889	EG10883	ECK3305	rplW	X	E	N	-
889	890	EG10884	ECK3296	rplX	X	E	N	-
890	891	EG10885	ECK2179	rplY	X	N	Y	4

891	892	EG10886	ECK3627	rpmB	X	E	N	-
892	893	EG10887	ECK3299	rpmC	X	E	N	-
893	894	EG10888	ECK3289	rpmD	X	E	N	-
894	895	EG10889	ECK3928	rpmE	X	N	Y	4
895	896	EG10890	ECK1075	rpmF	X	N	Y	4
896	897	EG10891	ECK3626	rpmG	X	N	Y	4
897	898	EG10892	ECK3695	rpmH	X	E	N	-
898	899	EG10893	ECK3282	rpoA	X	E	N	-
899	900	EG10894	ECK3978	rpoB	X	E	N	-
900	901	EG10895	ECK3979	rpoC	X	E	N	-
905	906	EG10900	ECK0902	rpsA	X	E	N	-
906	907	EG10901	ECK0168	rpsB	X	E	N	-
907	908	EG10902	ECK3301	rpsC	X	E	N	-
908	909	EG10903	ECK3283	rpsD	X	E	N	-
909	910	EG10904	ECK3290	rpsE	X	E	N	-
910	911	EG10905	ECK4196	rpsF	X	N	Y	4
911	912	EG10906	ECK3328	rpsG	X	E	N	-
912	913	EG10907	ECK3293	rpsH	X	E	N	-
913	914	EG10908	ECK3219	rpsI	X	E	N	-
914	915	EG10909	ECK3308	rpsJ	X	E	N	-
915	916	EG10910	ECK3284	rpsK	X	E	N	-
916	917	EG10911	ECK3329	rpsL	X	E	N	-
917	918	EG10912	ECK3285	rpsM	X	E	N	-
918	919	EG10913	ECK3294	rpsN	X	E	N	-
919	920	EG10914	ECK3154	rpsO	X	N	Y	4
920	921	EG10915	ECK2606	rpsP	X	E	N	-
921	922	EG10916	ECK3298	rpsQ	X	E	N	-
922	923	EG10917	ECK4198	rpsR	X	E	N	-
923	924	EG10918	ECK3303	rpsS	X	E	N	-
924	925	EG10919	ECK0024	rpsT	X	N	Y	4

925	926	EG10920	ECK3055	rpsU	X	N	Y	4
927	928	EG10922	ECK3966	coaA	Y	N	N	-
931	932	EG10928	ECK0372	sbmA	Y	N	N	-
932	933	EG10929	ECK3909	sbp	Y	N	N	-
933	934	EG10930	ECK1812	sdaA	Y	N	N	-
934	935	EG10931	ECK0712	sdhA	Y	N	N	-
935	936	EG10932	ECK0713	sdhB	Y	N	N	-
936	937	EG10933	ECK0710	sdhC	Y	N	N	-
937	938	EG10934	ECK0711	sdhD	Y	N	N	-
947	948	EG10944	ECK2909	serA	Y	N	N	-
948	949	EG10945	ECK4380	serB	Y	N	N	-
949	950	EG10946	ECK0898	serC	Y	N	N	-
950	951	EG10947	ECK0884	serS	Y	E	Y	1
951	952	EG10948	ECK1473	maeA/sfcA	Y	N	N	-
953	954	EG10950	ECK4384	slt	Y	N*	N	-
963	964	EG10960	ECK2932	speB	Y	N	N	-
964	965	EG10961	ECK2960	speC	Y	N*	N	-
966	967	EG10963	ECK0120	speE	Y	N	N	-
967	968	EG10964	ECK0681	speF	Y	N	N	-
968	969	EG10965	ECK3638	gmk	X	E	N	-
969	970	EG10966	ECK3640	spoT	Y	E	Y	1
974	975	EG10971	ECK2700	srlD	Y	N	N	-
976	977	EG10973	ECK2703	gutQ	Y	N	N	-
982	983	EG10979	ECK0714	sucA	Y	N	N	-
983	984	EG10980	ECK0715	sucB	Y	N	N	-
984	985	EG10981	ECK0716	sucC	Y	N	N	-
985	986	EG10982	ECK0717	sucD	Y	N	N	-
988	989	EG10985	ECK0054	surA	Y	N	N	-
993	994	EG10990	ECK3106	tdcB	Y	N	N	-
994	995	EG10991	ECK3105	tdcC	Y	N	N	-

996	997	EG10993	ECK3606	tdh	Y	N	N	-
997	998	EG10994	ECK1233	tdk	Y	N	N	-
1001	1002	EG10998	ECK0002	thrA	Y	N	N	-
1002	1003	EG10999	ECK0003	thrB	O	N	N!	-
1003	1004	EG11000	ECK0004	thrC	O	N	N!	-
1004	1005	EG11001	ECK1717	thrS	Y	E	Y	1
1005	1006	EG11002	ECK2823	thyA	X	N	Y	4
1006	1007	EG11003	ECK0430	tig	Y	N	N	-
1008	1009	EG11005	ECK3701	tnaA	Y	N*	N	-
1009	1010	EG11006	ECK3702	tnaB	Y	u	X	3
1011	1012	EG11008	ECK0729	tolB	Y	N	N	-
1012	1013	EG11009	ECK3026	tolC	Y	N	N	-
1018	1019	EG11015	ECK3911	tpiA	Y	N	N	-
1023	1024	EG11020	ECK1362	trkG	Y	N	N	-
1024	1025	EG11021	ECK3841	trkH	Y	N*	N	-
1026	1027	EG11023	ECK2604	trmD	Y	E	Y	1
1027	1028	EG11024	ECK1254	trpA	O	N	N!	-
1028	1029	EG11025	ECK1255	trpB	Y	N	N	-
1029	1030	EG11026	ECK1256	trpC	O	N	N!	-
1030	1031	EG11027	ECK1257	trpD	O	N	N!	-
1031	1032	EG11028	ECK1258	trpE	O	N	N!	-
1032	1033	EG11029	ECK4385	trpR	Y	N	N	-
1033	1034	EG11030	ECK3371	trpS	Y	E	Y	1
1035	1036	EG11032	ECK0879	trxB	O	N	N!	-
1038	1039	EG11035	ECK0405	tsx	Y	N	N	-
1042	1043	EG11039	ECK2597	tyrA	O	N	N!	-
1043	1044	EG11040	ECK4046	tyrB	Y	N	N	-
1044	1045	EG11041	ECK1906	tyrP	Y	N	N	-
1045	1046	EG11042	ECK1319	tyrR	Y	N	N	-
1046	1047	EG11043	ECK1633	tyrS	Y	E	Y	1

1048	1049	EG11045	ECK3825	udp	Y	N	N	-
1049	1050	EG11046	ECK3436	ugpA	Y	N	N	-
1050	1051	EG11047	ECK3437	ugpB	Y	N	N	-
1051	1052	EG11048	ECK3434	ugpC	Y	N	N	-
1052	1053	EG11049	ECK3435	ugpE	Y	N	N	-
1053	1054	EG11050	ECK3433	ugpQ	Y	N	N	-
1063	1064	EG11060	ECK0474	ushA	Y	N	N	-
1068	1069	EG11065	ECK1514	uxaB	Y	N*	N	-
1069	1070	EG11066	ECK4313	uxuA	Y	N	N	-
1070	1071	EG11067	ECK4251	valS	Y	E	Y	1
1077	1078	EG11074	ECK3554	xylA	Y	N	N	-
1078	1079	EG11075	ECK3553	xylB	Y	N	N	-
1079	1080	EG11076	ECK4023	xylE	Y	N	N	-
1081	1082	EG11079	ECK0026	ribF	Y	E	Y	1
1082	1083	EG11080	ECK0029	fkpB	Y	N	N	-
1083	1084	EG11081	ECK0030	ispH	Y	E	Y	1
1084	1085	EG11082	ECK0031	rihC	Y	N	N	-
1092	1093	EG11090	ECK0158	mtn/pfs	Y	N	N	-
1097	1098	EG11095	ECK0398	acpH/yajB	Y	N	N	-
1104	1105	EG11102	ECK0471	gsk	Y	N	N	-
1112	1113	EG11112	ECK0876	aat	Y	N	N	-
1133	1134	EG11135	ECK1766	pncA	Y	N*	N	-
1136	1137	EG11138	ECK1866	nudB	Y	N	N	-
1141	1142	EG11143	ECK2224	ubiG	Y	N	N	-
1155	1156	EG11158	ECK2907	fau/ygfA	Y	N*	N	-
1164	1165	EG11167	ECK3031	zupT/ygiE	Y	N	N	-
1165	1166	EG11168	ECK3051	ttdA	Y	N	N	-
1166	1167	EG11169	ECK3052	ttdB	Y	N	N	-
1168	1169	EG11172	ECK3104	tdcD	Y	N	N	-
1171	1172	EG11175	ECK3112	garK	Y	N*	N	-

1172	1173	EG11176	ECK3113	garR	Y	N	N	-
1173	1174	EG11177	ECK3155	truB	Y	N	N	-
1184	1185	EG11189	ECK3611	waaC/rfaC	Y	N	N	-
1185	1186	EG11190	ECK3624	coaD	Y	E	Y	1
1189	1190	EG11194	ECK3644	xanP/yicE	Y	N	N	-
1190	1191	EG11195	ECK3689	yidA	Y	N	N	-
1193	1194	EG11198	ECK3723	glmU	Y	E	Y	1
1196	1197	EG11202	ECK3807	yigB	Y	N	N	-
1198	1199	EG11204	ECK3959	murI	Y	E	Y	1
1199	1200	EG11205	ECK3964	murB	Y	E	Y	1
1213	1214	EG11221	ECK1853	zwf	Y	N	N	-
1215	1216	EG11223	ECK3951	argH	X	N	Y	4
1217	1218	EG11225	ECK4132	dcuA	Y	N	N	-
1218	1219	EG11226	ECK0076	leuA	O	N	N!	-
1219	1220	EG11227	ECK1468	fdnG	Y	N	N	-
1220	1221	EG11228	ECK1469	fdnH	Y	N	N	-
1221	1222	EG11229	ECK1470	fdnI	Y	N	N	-
1223	1224	EG11231	ECK1715	rpmI	X	N*	Y	4
1224	1225	EG11232	ECK3286	rpmJ	X	N	Y	4
1226	1227	EG11234	ECK1690	ydiB	Y	N	N	-
1231	1232	EG11239	ECK0755	ybhA	Y	N	N	-
1235	1236	EG11245	ECK2926	yggF	Y	N	N	-
1247	1248	EG11258	ECK0895	focA	Y	N	N	-
1248	1249	EG11259	ECK4175	rnr	Y	N*	N	-
1253	1254	EG11265	ECK0901	cmk	Y	N	N	-
1256	1257	EG11268	ECK3274	fmt	Y	E	Y	1
1269	1270	EG11283	ECK4096	phnE	Y	N	N	-
1270	1271	EG11284	ECK0179	fabZ	Y	E	Y	1
1272	1273	EG11286	ECK3948	argE	X	N	Y	4
1274	1275	EG11288	ECK0389	mak	Y	N	N	-

1275	1276	EG11289	ECK2140	preA/yeiA	Y	N	N	-
1278	1279	EG11292	ECK2997	yghA	Y	N	N	-
1280	1281	EG11294	ECK1196	ispE	Y	E	Y	1
1284	1285	EG11298	ECK3236	yhdE	Y	N	N	-
1285	1286	EG11299	ECK3235	rng	Y	N	N	-
1292	1293	EG11306	ECK0621	lipA	Y	N	N	-
1297	1298	EG11313	ECK3900	rhaT	Y	N	N	-
1298	1299	EG11314	ECK1117	purB	O	E	Y!	2
1299	1300	EG11315	ECK3241	yhdH	Y	N	N	-
1301	1302	EG11317	ECK1078	fabD	Y	E	Y	1
1302	1303	EG11318	ECK1079	fabG	Y	E	Y	1
1305	1306	EG11321	ECK0408	ribD	Y	E	Y	1
1306	1307	EG11322	ECK0409	ribE	Y	E	Y	1
1308	1309	EG11324	ECK2902	ubiH	Y	N	N	-
1309	1310	EG11325	ECK1271	acnA	Y	N	N	-
1310	1311	EG11326	ECK0334	codA	Y	N	N	-
1313	1314	EG11329	ECK2655	gabD	Y	N	N	-
1314	1315	EG11330	ECK2657	gabP	Y	N	N	-
1315	1316	EG11331	ECK1272	ribA	Y	E	Y	1
1316	1317	EG11332	ECK2494	upp	Y	N	N	-
1317	1318	EG11333	ECK2901	ubiI/visC	Y	N	N	-
1319	1320	EG11335	ECK4382	nadR	Y	N	N	-
1321	1322	EG11337	ECK2149	lysP	Y	N	N	-
1323	1324	EG11339	ECK3621	waaG/rfaG	Y	N	N	-
1324	1325	EG11340	ECK3620	waaP/rfaP	Y	N*	N	-
1325	1326	EG11341	ECK3622	waaQ/rfaQ	Y	N	N	-
1333	1334	EG11351	ECK3618	waaB/rfaB	Y	N	N	-
1334	1335	EG11352	ECK3617	waaO/rfaI	Y	N	N	-
1335	1336	EG11353	ECK3616	waaJ/rfaJ	Y	N	N	-
1340	1341	EG11358	ECK3178	murA	Y	E	Y	1

1349	1350	EG11368	ECK2256	menB	Y	N	N	-
1350	1351	EG11369	ECK4031	ubiC	Y	N*	N	-
1351	1352	EG11370	ECK4032	ubiA	Y	E	Y	1
1352	1353	EG11371	ECK2558	pgpC/yfhB	Y	N*	N	-
1353	1354	EG11372	ECK2557	tadA	Y	E	Y	1
1354	1355	EG11373	ECK2556	mltF/yfhD	Y	N	N	-
1355	1356	EG11374	ECK0141	folK	Y	E	Y	1
1356	1357	EG11375	ECK2146	folE	Y	E	Y	1
1358	1359	EG11377	ECK3009	plsC	Y	E	Y	1
1360	1361	EG11379	ECK0970	appB	Y	N	N	-
1361	1362	EG11380	ECK0969	appC	Y	N	N	-
1372	1373	EG11392	ECK1174	nhaB	Y	N	N	-
1376	1377	EG11396	ECK3835	ubiD	Y	E	Y	1
1381	1382	EG11402	ECK4234	treC	Y	N	N	-
1384	1385	EG11405	ECK0878	cydD	Y	N	N	-
1385	1386	EG11406	ECK1658	ribC	Y	E	Y	1
1386	1387	EG11407	ECK1177	dadA	Y	N	N	-
1387	1388	EG11408	ECK1178	dadX	Y	N	N	-
1388	1389	EG11409	ECK0906	lpxK	Y	E	Y	1
1395	1396	EG11417	ECK4233	nrdD	Y	N	N	-
1396	1397	EG11418	ECK2059	dcd	Y	N	N	-
1400	1401	EG11423	ECK3613	waaU/rfaK	Y	E	Y	1
1402	1403	EG11425	ECK3615	waaY/rfaY	Y	N*	N	-
1403	1404	EG11426	ECK3614	waaZ/rfaZ	Y	N*	N	-
1404	1405	EG11427	ECK2930	tktA	Y	N*	N	-
1405	1406	EG11428	ECK3954	sthA	Y	N	N	-
1410	1411	EG11433	ECK4215	msrA	Y	N	N	-
1413	1414	EG11436	ECK4262	ahr/yjgB	Y	N*	N	-
1417	1418	EG11440	ECK3273	def	Y	E	Y	1
1419	1420	EG11442	ECK2900	gcvT	Y	N	N	-

1420	1421	EG11443	ECK2910	rpiA	Y	N	N	-
1421	1422	EG11444	ECK2913	scpA/yliK	Y	N	N	-
1424	1425	EG11448	ECK4062	acs	Y	N	N	-
1427	1428	EG11451	ECK3778	wecB/rffE	Y	N*	N	-
1429	1430	EG11453	ECK3780	rffG	Y	N*	N	-
1430	1431	EG11454	ECK3781	rffH	Y	N	N	-
1432	1433	EG11456	ECK3783	wecE/rffA	Y	N	N	-
1438	1439	EG11463	ECK3810	corA	Y	N	N	-
1443	1444	EG11468	ECK3817	rhtC	Y	N*	N	-
1445	1446	EG11470	ECK3820	yigL	Y	N*	N	-
1448	1449	EG11473	ECK3827	ubiE	Y	N*	N	-
1454	1455	EG11485	ECK3842	hemG	Y	E	Y	1
1456	1457	EG11487	ECK1634	pdxH	Y	N	N	-
1459	1460	EG11490	ECK1487	gadB	Y	N	N	-
1461	1462	EG11492	ECK4043	qorA/qor	Y	N	N	-
1462	1463	EG11493	ECK1082	pabC	Y	N	N	-
1470	1471	EG11501	ECK4110	adiA	Y	N	N	-
1473	1474	EG11504	ECK0197	metQ	Y	N	N	-
1477	1478	EG11508	ECK1474	sra	X	N*	Y	4
1479	1480	EG11510	ECK2497	ppk	Y	N	N	-
1481	1482	EG11512	ECK0010	satP/yaaH	Y	N	N	-
1485	1486	EG11517	ECK3917	glpX	Y	N	N	-
1495	1496	EG11528	ECK1283	fabI	Y	E	Y	1
1497	1498	EG11530	ECK1803	fadD	Y	N	N	-
1499	1500	EG11532	ECK2255	menC	Y	N	N	-
1506	1507	EG11539	ECK0170	pyrH	O	E	Y!	2
1507	1508	EG11540	ECK0995	wrbA	Y	N	N	-
1508	1509	EG11541	ECK3741	kup/trkD	Y	N*	N	-
1509	1510	EG11542	ECK0488	tesA	Y	N	N	-
1510	1511	EG11543	ECK3989	hemE	Y	N	N	-

1513	1514	EG11546	ECK0108	nadC	Y	N	N	-
1514	1515	EG11547	ECK1648	rnt	Y	N	N	-
1517	1518	EG11551	ECK2707	hypF	X	N	Y	4
1519	1520	EG11553	ECK3165	glmM	X	N	Y	4
1522	1523	EG11556	ECK0008	talB	Y	N	N	-
1533	1534	EG11568	ECK0047	kefF	Y	N	N	-
1534	1535	EG11569	ECK0055	lptD/imp	Y	E	Y	1
1540	1541	EG11575	ECK0073	leuD	O	N	N!	-
1541	1542	EG11576	ECK0074	leuC	X	N	Y	4
1542	1543	EG11577	ECK0075	leuB	Y	N*	N	-
1547	1548	EG11583	ECK0500	gcl	Y	N	N	-
1548	1549	EG11585	ECK3986	thiC	Y	N	N	-
1549	1550	EG11586	ECK3985	thiE	Y	N	N	-
1550	1551	EG11587	ECK3984	thiF	Y	N*	N	-
1551	1552	EG11589	ECK3982	thiG	Y	N*	N	-
1552	1553	EG11590	ECK3981	thiH	Y	N	N	-
1553	1554	EG11591	ECK0623	lipB	Y	N	N	-
1564	1565	EG11603	ECK3044	ygiF	Y	N	N	-
1567	1568	EG11606	ECK1244	kch	Y	N	N	-
1569	1570	EG11608	ECK1243	clsA/cls	Y	N	N	-
1573	1574	EG11612	ECK4104	proP	Y	N	N	-
1576	1577	EG11615	ECK4106	basR	Y	N	N	-
1579	1580	EG11618	ECK2075	baeR	Y	N	N	-
1581	1582	EG11620	ECK1281	rnb	Y	N	N	-
1582	1583	EG11621	ECK0199	metN	Y	N	N	-
1584	1585	EG11623	ECK2792	sdaB	Y	N	N	-
1585	1586	EG11624	ECK0855	artP	Y	N	N	-
1587	1588	EG11626	ECK0853	artQ	Y	N	N	-
1588	1589	EG11627	ECK0852	artM	Y	N	N	-
1589	1590	EG11628	ECK0851	artJ	Y	N	N	-

1590	1591	EG11629	ECK0845	potF	Y	N	N	-
1591	1592	EG11630	ECK0846	potG	Y	N*	N	-
1592	1593	EG11631	ECK0847	potH	Y	N	N	-
1593	1594	EG11632	ECK0848	potI	Y	N	N	-
1596	1597	EG11636	ECK1521	ydeA	Y	N	N	-
1598	1599	EG11639	ECK1526	eamA	Y	N*	N	-
1604	1605	EG11647	ECK0184	accA	Y	E	Y	1
1605	1606	EG11648	ECK0207	dkgB	Y	N	N	-
1619	1620	EG11663	ECK3336	slyD	Y	N	N	-
1625	1626	EG11669	ECK2214	atoD	Y	N	N	-
1626	1627	EG11670	ECK2215	atoA	Y	N	N	-
1628	1629	EG11672	ECK2217	atoB	Y	N	N	-
1629	1630	EG11673	ECK3048	folB	Y	N	N	-
1631	1632	EG11675	ECK0133	panB	Y	N	N	-
1635	1636	EG11679	ECK2832	aas	Y	N	N	-
1636	1637	EG11680	ECK3190	lptB/yhbG	Y	N	N	-
1645	1646	EG11689	ECK3653	nepI/yicM	Y	N	N	-
1648	1649	EG11692	ECK3656	adeD/ade	X	N	Y	4
1654	1655	EG11698	ECK0745	galM	Y	N	N	-
1655	1656	EG11699	ECK0744	gpmA	Y	N	N	-
1657	1658	EG11701	ECK2060	udk	Y	N	N	-
1659	1660	EG11703	ECK0457	acrA	Y	N	N	-
1660	1661	EG11704	ECK0456	acrB	Y	N	N	-
1675	1676	EG11723	ECK3706	yieF	Y	N	N	-
1687	1688	EG11735	ECK1009	efeB/ycdB	Y	N	N	-
1688	1689	EG11736	ECK0200	gmhB	Y	N	N	-
1689	1690	EG11737	ECK0198	metI	Y	N	N	-
1697	1698	EG11746	ECK0132	panC	Y	N	N	-
1702	1703	EG11751	ECK1895	otsA	Y	N	N	-
1703	1704	EG11752	ECK1896	otsB	Y	N	N	-

1704	1705	EG11753	ECK1210	chaA	Y	N	N	-
1705	1706	EG11754	ECK0072	setA	Y	N*	N	-
1707	1708	EG11758	ECK4163	nnr/yjeF	Y	N	N	-
1720	1721	EG11773	ECK2271	nuoM	Y	N	N	-
1721	1722	EG11774	ECK2278	nuoF	Y	N	N	-
1743	1744	EG11796	ECK4378	lplA	Y	N	N	-
1744	1745	EG11797	ECK2459	talA	Y	N	N	-
1756	1757	EG11809	ECK1850	purT	Y	N	N	-
1757	1758	EG11810	ECK2898	gcvP	Y	N	N	-
1758	1759	EG11811	ECK4183	aidB	Y	N*	N	-
1762	1763	EG11816	ECK2741	ispF	Y	E	Y	1
1763	1764	EG11817	ECK2739	umpG/surE	Y	N	N	-
1764	1765	EG11819	ECK1227	purU	Y	N	N	-
1766	1767	EG11821	ECK0445	amtB	Y	N	N	-
1767	1768	EG11822	ECK1296	puuB	Y	N	N	-
1771	1772	EG11826	ECK0972	etk/yccC	Y	N	N	-
1772	1773	EG11827	ECK4083	rpiB	Y	N	N	-
1776	1777	EG11831	ECK3851	srkA/yihE	Y	N	N	-
1780	1781	EG11836	ECK3860	hemN	Y	N	N	-
1789	1790	EG11845	ECK3873	yihS	Y	N*	N	-
1791	1792	EG11847	ECK3875	yihU	Y	N	N	-
1792	1793	EG11848	ECK3876	yihV	Y	N	N	-
1794	1795	EG11850	ECK3878	yihX	Y	N*	N	-
1800	1801	EG11856	ECK3885	fdoI	Y	N	N	-
1801	1802	EG11857	ECK3886	fdoH	Y	N	N	-
1802	1803	EG11858	ECK3887	fdoG	Y	N	N	-
1809	1810	EG11865	ECK3894	rhaM/yiiL	Y	N	N	-
1810	1811	EG11866	ECK3895	rhaD	Y	u	X	3
1811	1812	EG11867	ECK3896	rhaA	Y	N*	N	-
1812	1813	EG11868	ECK3897	rhaB	Y	u	X	3

1816	1817	EG11873	ECK3907	fieF	Y	N	N	-
1823	1824	EG11880	ECK3922	menA	Y	N	N	-
1839	1840	EG11896	ECK2008	plaP/yeeF	Y	N	N	-
1844	1845	EG11902	ECK3935	yijE	Y	N*	N	-
1846	1847	EG11904	ECK3937	gldA	Y	N*	N	-
1860	1861	EG11919	ECK4012	yjbB	Y	N	N	-
1875	1876	EG11934	ECK4047	aphA	Y	N	N	-
1880	1881	EG11939	ECK4057	ghxP/yjcD	Y	N	N	-
1896	1897	EG11956	ECK4077	alsK	Y	E	Y	1
1897	1898	EG11957	ECK4078	alsE	Y	N	N	-
1898	1899	EG11958	ECK4079	alsC	Y	N	N	-
1899	1900	EG11959	ECK4080	alsA	Y	N	N	-
1900	1901	EG11961	ECK3593	lldP	Y	N	N	-
1902	1903	EG11963	ECK3595	lldD	Y	N	N	-
1909	1910	EG11971	ECK2388	nupC	Y	N	N	-
1916	1917	EG11978	ECK2033	rfbA	Y	N	N	-
1917	1918	EG11979	ECK2032	rfbC	Y	N	N	-
1919	1920	EG11981	ECK2030	glf	Y	N	N	-
1921	1922	EG11983	ECK2028	wbbI	Y	N*	N	-
1956	1957	EG12026	ECK2147	yeiG	Y	N	N	-
1960	1961	EG12030	ECK2155	rihB	Y	N	N	-
1963	1964	EG12033	ECK2158	psuG/yeiN	Y	N	N	-
1964	1965	EG12034	ECK2163	setB	Y	N	N	-
1967	1968	EG12037	ECK2171	yejA	Y	N*	N	-
1968	1969	EG12038	ECK2172	yejB	Y	N	N	-
1969	1970	EG12040	ECK2173	yejE	Y	N	N	-
1970	1971	EG12041	ECK2174	yejF	Y	N	N	-
1991	1992	EG12069	ECK2202	mqo	Y	N	N	-
1994	1995	EG12075	ECK3460	nikA	Y	N	N	-
1995	1996	EG12076	ECK3461	nikB	Y	N	N	-

1996	1997	EG12077	ECK3462	nikC	Y	N	N	-
1997	1998	EG12078	ECK3463	nikD	Y	N	N	-
1998	1999	EG12079	ECK3464	nikE	Y	N	N	-
2001	2002	EG12082	ECK2282	nuoA	Y	N*	N	-
2002	2003	EG12083	ECK2281	nuoB	Y	N	N	-
2003	2004	EG12084	ECK2280	nuoC	Y	N	N	-
2004	2005	EG12086	ECK2279	nuoE	Y	N	N	-
2005	2006	EG12087	ECK2277	nuoG	Y	N*	N	-
2006	2007	EG12088	ECK2276	nuoH	Y	N	N	-
2007	2008	EG12089	ECK2275	nuoI	Y	N	N	-
2008	2009	EG12090	ECK2274	nuoJ	Y	N	N	-
2009	2010	EG12091	ECK2273	nuoK	Y	N	N	-
2010	2011	EG12092	ECK2272	nuoL	Y	N	N	-
2011	2012	EG12093	ECK2270	nuoN	Y	N	N	-
2018	2019	EG12100	ECK2460	tktB	Y	N	N	-
2020	2021	EG12102	ECK3396	feoB	Y	N	N	-
2032	2033	EG12115	ECK4365	yjjG	Y	N	N	-
2040	2041	EG12124	ECK2303	hisJ	Y	N	N	-
2041	2042	EG12125	ECK2302	hisQ	Y	N	N	-
2044	2045	EG12129	ECK2493	uraA	Y	N	N	-
2049	2050	EG12134	ECK0802	rhtA	Y	N	N	-
2056	2057	EG12142	ECK2791	sdaC	Y	N	N	-
2058	2059	EG12144	ECK0676	pgm	Y	N	N	-
2061	2062	EG12148	ECK2938	galP	X	N	Y	4
2064	2065	EG12152	ECK4261	idnK	Y	N	N	-
2077	2078	EG12168	ECK0395	brnQ	Y	N	N	-
2091	2092	EG12189	ECK2431	hemF	Y	N	N	-
2094	2095	EG12192	ECK2611	nadK/yfjB	Y	E	Y	1
2110	2111	EG12210	ECK3610	waaF/rfaF	Y	N	N	-
2115	2116	EG12215	ECK3453	zntA	Y	N	N	-

2120	2121	EG12221	ECK3459	acpT	Y	N	N	-
2128	2129	EG12230	ECK3478	pitA	Y	N	N	-
2130	2131	EG12232	ECK3481	dtpB/yhiP	Y	N	N	-
2134	2135	EG12236	ECK3487	arsB	Y	N*	N	-
2135	2136	EG12237	ECK3488	arsC	Y	N	N	-
2142	2143	EG12245	ECK3504	treF	Y	N	N	-
2149	2150	EG12253	ECK3511	kdgK	Y	N	N	-
2162	2163	EG12267	ECK3533	eptB	Y	N	N	-
2167	2168	EG12272	ECK3540	ghrB/tiaE	Y	N	N	-
2170	2171	EG12275	ECK3556	xylG	Y	N	N	-
2171	2172	EG12276	ECK3557	xylH	Y	N	N	-
2174	2175	EG12279	ECK3564	yiaK	Y	N	N	-
2179	2180	EG12284	ECK3569	lyxK	Y	N	N	-
2180	2181	EG12285	ECK3570	sgbH	Y	N	N	-
2181	2182	EG12286	ECK3571	sgbU	Y	N	N	-
2182	2183	EG12287	ECK3572	sgbE	Y	N	N	-
2187	2188	EG12292	ECK3577	aldB	Y	N*	N	-
2188	2189	EG12293	ECK3578	yiaY	Y	N*	N	-
2191	2192	EG12296	ECK3602	gpmM	Y	N	N	-
2193	2194	EG12298	ECK3604	yibQ	Y	N*	N	-
2195	2196	EG12302	ECK1084	tmk	X	E	N	-
2201	2202	EG12310	ECK2520	pepB	Y	N*	N	-
2203	2204	EG12312	ECK0103	coaE	Y	N	N	-
2207	2208	EG12316	ECK0117	acnB	Y	N	N	-
2208	2209	EG12318	ECK0122	cueO	Y	N	N	-
2209	2210	EG12319	ECK0125	can	Y	E	Y	1
2220	2221	EG12331	ECK0154	clcA	Y	N	N	-
2234	2235	EG12345	ECK1532	ydfG	Y	N	N	-
2235	2236	EG12347	ECK1916	tcyN/yecC	Y	N	N	-
2236	2237	EG12352	ECK3767	ppiC	Y	N	N	-

2239	2240	EG12357	ECK1699	fadK/ydiD	Y	N*	N	-
2240	2241	EG12358	ECK0924	ssuB	Y	N	N	-
2243	2244	EG12361	ECK2840	kduD	Y	N	N	-
2244	2245	EG12362	ECK2259	menF	Y	N*	N	-
2250	2251	EG12368	ECK1860	znuB	Y	N	N	-
2262	2263	EG12381	ECK2670	nrdF	Y	N	N	-
2263	2264	EG12384	ECK0514	ybcF	Y	N	N	-
2266	2267	EG12387	ECK2654	lhgD/ygaF	Y	N	N	-
2269	2270	EG12392	ECK1194	dauA/ychM	Y	N	N	-
2271	2272	EG12394	ECK1776	msrB/yeaA	Y	N	N	-
2275	2276	EG12399	ECK1186	dhaM/dhaH	Y	N	N	-
2285	2286	EG12410	ECK2462	nudK/yffH	Y	N	N	-
2286	2287	EG12411	ECK2034	rfbD	Y	N	N	-
2287	2288	EG12412	ECK2035	rfbB	Y	N	N	-
2292	2293	EG12419	ECK2089	gatY	Y	N	N	-
2296	2297	EG12424	ECK1200	prmC	Y	E	Y	1
2301	2302	EG12432	ECK1692	ydiF	Y	N	N	-
2305	2306	EG12437	ECK2254	menE	Y	N	N	-
2306	2307	EG12438	ECK2257	menH/yfbB	Y	N	N	-
2308	2309	EG12440	ECK4228	mpl	Y	N	N	-
2312	2313	EG12445	ECK2576	eamB/yfiK	Y	N	N	-
2319	2320	EG12458	ECK4081	alsB	Y	N	N	-
2321	2322	EG12469	ECK4124	dtpC/yjdL	Y	N	N	-
2325	2326	EG12504	ECK4204	cycA	Y	N	N	-
2330	2331	EG12525	ECK4237	mgtA	Y	N	N	-
2337	2338	EG12563	ECK4312	gntP	Y	N	N	-
2338	2339	EG12576	ECK4328	mdtM/yjiO	Y	N	N	-
2340	2341	EG12591	ECK4349	opgB/mdoB	Y	N	N	-
2341	2342	EG12600	ECK4386	yjjX	Y	N	N	-
2343	2344	EG12606	ECK1081	fabF	Y	N	N	-

2351	2352	EG12625	ECK3530	dppB	Y	N	N	-
2352	2353	EG12626	ECK3529	dppC	Y	N	N	-
2353	2354	EG12627	ECK3528	dppD	Y	N	N	-
2354	2355	EG12628	ECK3527	dppF	Y	N	N	-
2355	2356	EG12629	ECK3422	gntK	Y	N	N	-
2358	2359	EG12633	ECK3025	nudF	X	N	Y	4
2360	2361	EG12656	ECK3019	mdaB	X	N	Y	4
2363	2364	EG12661	ECK0647	gltJ	Y	N	N	-
2364	2365	EG12662	ECK0646	gltK	Y	N	N	-
2365	2366	EG12663	ECK0645	gltL	Y	N	N	-
2366	2367	EG12666	ECK0517	lpxH	Y	E	Y	1
2372	2373	EG12693	ECK1811	nudL/yeaB	X	N	Y	4
2383	2384	EG12712	ECK2681	luxS	Y	N	N	-
2384	2385	EG12713	ECK1467	yddG	Y	N*	N	-
2385	2386	EG12715	ECK0172	dxr	Y	E	Y	1
2388	2389	EG12732	ECK3080	sstT	Y	N	N	-
2389	2390	EG12734	ECK3082	uxaA	Y	N	N	-
2390	2391	EG12738	ECK3084	exuT	Y	N	N	-
2395	2396	EG12768	ECK3125	kbaY	Y	N	N	-
2435	2436	EG12882	ECK2982	gss	Y	N	N	-
2436	2437	EG12883	ECK2981	pitB	Y	N	N	-
2440	2441	EG12957	ECK2384	glk	Y	N	N	-
2442	2443	EG13159	ECK2400	xapB	Y	N	N	-
2449	2450	EG13270	ECK0866	aqpZ	Y	N	N	-
2451	2452	EG20044	ECK3513	dctA	Y	N	N	-
2452	2453	EG20049	ECK3685	dgoA	Y	N*	N	-
2456	2457	EG20091	ECK3598	gpsA	Y	E	Y	1
2457	2458	EG20098	ECK0124	hpt	Y	N*	N	-
2458	2459	EG20110	ECK3338	kefB	X	N	Y	4
2459	2460	EG20173	ECK2291	pta	Y	N	N	-

2460	2461	EG20248	ECK4314	uxuB	Y	N	N	-
2462	2463	EG20252	ECK3555	xylF	Y	N	N	-
2464	2465	EG20257	ECK2669	nrdE	X	N	Y	4
2465	2466	EG50001	ECK3175	rplU	X	E	N	-
2466	2467	EG50002	ECK3174	rpmA	X	E	N	-
2470	2471	EG50006	ECK2436	eutB	X	N*	Y	4
2471	2472	EG50007	ECK2435	eutC	Y	N	N	-
2473	2474	EG50009	ECK3502	gadA	X	N	Y	4
2474	2475	EG50010	ECK0353	frmA	X	N	Y	4
2475	2476	EG50011	ECK3166	folP	X	N*	Y	4
2488	2489	G0-10439	ECK0687	kdpF	Y	N*	Y	-
2490	2491	G0-10441	ECK0723	cydX/ybgT	X	N	Y	4
2599	2600	G0-8601	ECK2972	glcF	X	N*	Y	4
2644	2645	G436	ECK3213	nanT	X	N*	Y	4
2646	2647	G58	ECK0453	maa	X	N	Y	4
2647	2648	G592	ECK1377	ldhA	Y	N	N	-
2655	2656	G6092	ECK0173	uppS/ispU	X	E	N	-
2657	2658	G6094	ECK0185	ldcC	X	N	Y	4
2659	2660	G6096	ECK0187	tilS	Y	E	Y	1
2662	2663	G6099	ECK0212	gloB	X	N	Y	4
2668	2669	G6105	ECK0222	fadE	Y	N	N	-
2669	2670	G6106	ECK0223	gmhA/lpcA	Y	N	N	-
2699	2700	G6136	ECK0263	mmuM	Y	N	N	-
2703	2704	G6140	ECK0269	yagE	X	N	Y	4
2704	2705	G6141	ECK0270	yagF	X	N	Y	4
2749	2750	G6190	ECK0323	yahK	Y	N	N	-
2755	2756	G6196	ECK0329	prpB	Y	N	N	-
2756	2757	G6198	ECK0330	prpC	Y	N	N	-

2757	2758	G6199	ECK0331	prpD	Y	N	N	-
2758	2759	G6200	ECK0332	prpE	Y	N	N	-
2760	2761	G6205	ECK0349	mhpE	Y	N	N	-
2763	2764	G6208	ECK0352	frmB	Y	N	N	-
2785	2786	G6234	ECK0411	thiL	Y	E	Y	1
2787	2788	G6237	ECK0414	dxs	Y	E	Y	1
2788	2789	G6238	ECK0417	thiI	Y	N	N	-
2789	2790	G6239	ECK0419	panE	Y	N	N	-
2794	2795	G6244	ECK0437	fadM/ybaW	Y	N	N	-
2796	2797	G6246	ECK0440	cof	Y	N	N	-
2807	2808	G6260	ECK0478	copA	X	N	Y	4
2808	2809	G6261	ECK0479	glsA/ybaS	X	N	Y	4
2816	2817	G6269	ECK0487	ybbO	X	N	Y	4
2823	2824	G6277	ECK0501	hyi	Y	N	N	-
2824	2825	G6278	ECK0502	glxR	X	N	Y	4
2826	2827	G6281	ECK0505	allB	X	N	Y	4
2828	2829	G6283	ECK0507	glxK	Y	N	N	-
2829	2830	G6284	ECK0508	allE/ylbA	Y	N	N	-
2830	2831	G6285	ECK0509	allC	Y	N	N	-
2852	2853	G6310	ECK0546	rrrD/ybcS	Y	N	N	-
2870	2871	G6329	ECK0594	ybdL	Y	N	N	-
2884	2885	G6344	ECK0611	citC	Y	N*	N	-
2890	2891	G6350	ECK0632	nadD	Y	E	Y	1
2898	2899	G6358	ECK0644	rihA	Y	N	N	-
2899	2900	G6359	ECK0648	gltI	Y	N	N	-
2904	2905	G6364	ECK0653	miaB	Y	N	N	-
2905	2906	G6365	ECK0654	ubiF	Y	N	N	-
2911	2912	G6378	ECK0698	dtpD/ybgH	Y	N	N	-
2921	2922	G6393	ECK0741	zitB	Y	N	N	-
2925	2926	G6397	ECK0756	pgl/ybhE	Y	N	N	-

2934	2935	G6406	ECK0778	clsB/ybhO	Y	N	N	-
2952	2953	G6425	ECK0812	ybiV	Y	N	N	-
2956	2957	G6429	ECK0819	gsiA/yliA	Y	N	N	-
2957	2958	G6430	ECK0820	gsiB/yliB	Y	N	N	-
2958	2959	G6431	ECK0821	gsiC/yliC	Y	N	N	-
2959	2960	G6432	ECK0822	gsiD/yliD	Y	N	N	-
2964	2965	G6437	ECK0827	yliI	X	N	Y	4
2966	2967	G6439	ECK0831	ybjG	Y	N	N	-
2967	2968	G6440	ECK0832	mdfA/cmr	X	N	Y	4
2969	2970	G6442	ECK0834	ybjI	Y	N*	N	-
2982	2983	G6455	ECK0861	ltaE	Y	N	N	-
2985	2986	G6458	ECK0865	lysO/ybjE	Y	N	N	-
2995	2996	G6468	ECK0896	ycaO	Y	N*	N	-
3002	3003	G6475	ECK0918	gloC/ycbL	Y	N	N	-
3003	3004	G6476	ECK0925	ssuC	Y	N*	N	-
3004	3005	G6477	ECK0926	ssuD	Y	N	N	-
3005	3006	G6478	ECK0927	ssuA	X	N*	Y	4
3006	3007	G6479	ECK0928	ssuE	Y	N	N	-
3024	3025	G6497	ECK0954	mgsA	X	N	Y	4
3030	3031	G6503	ECK0973	etp	Y	N*	N	-
3043	3044	G6517	ECK0997	rutG/ycdG	X	N	Y	4
3054	3055	G6530	ECK1013	pgaB/ycdR	Y	N	N	-
3060	3061	G6539	ECK1019	ghrA/ycdW	X	N	Y	4
3072	3073	G6551	ECK1032	clsC/ymdC	Y	N	N	-
3081	3082	G6561	ECK1054	murJ/mviN	Y	E	Y	1
3096	3097	G6576	ECK1105	nagK/ycfX	Y	N	N	-
3100	3101	G6580	ECK1120	nudJ/ymfB	Y	N	N	-
3141	3142	G6622	ECK1181	emtA	X	N	Y	4
3145	3146	G6626	ECK1187	dhaL	Y	N	N	-
3146	3147	G6627	ECK1188	dhaK	X	N*	Y	4

3162	3163	G6643	ECK1291	puuP	X	N*	Y	4
3163	3164	G6644	ECK1292	puuA	X	N*	Y	4
3164	3165	G6645	ECK1293	puuD	Y	N*	N	-
3165	3166	G6646	ECK1297	puuE	Y	N	N	-
3174	3175	G6655	ECK1312	ycjU	X	N	Y	4
3180	3181	G6661	ECK1321	ycjG	X	N*	Y	4
3181	3182	G6662	ECK1322	mpaA	X	N	Y	4
3184	3185	G6665	ECK1326	mppA	Y	N	N	-
3217	3218	G6701	ECK1374	ydbK	Y	N	N	-
3223	3224	G6708	ECK1384	paaZ/maoC	X	N	Y	4
3224	3225	G6709	ECK1385	paaA	Y	N	N	-
3225	3226	G6710	ECK1386	paaB	Y	N	N	-
3226	3227	G6711	ECK1387	paaC	X	N	Y	4
3228	3229	G6713	ECK1389	paaE	Y	N	N	-
3229	3230	G6714	ECK1390	paaF	Y	N	N	-
3230	3231	G6715	ECK1391	paaG	X	N	Y	4
3231	3232	G6716	ECK1392	paaH	X	N	Y	4
3232	3233	G6717	ECK1393	paaI	X	N	Y	4
3233	3234	G6718	ECK1394	paaJ	X	N	Y	4
3234	3235	G6719	ECK1395	paaK	X	N	Y	4
3244	3245	G6731	ECK1405	azoR	X	N	Y	4
3262	3263	G6755	ECK1438	patD/ydcW	X	N	Y	4
3266	3267	G6759	ECK1442	mnaT/yncA	Y	N	N	-
3271	3272	G6764	ECK1447	ansP	X	N*	Y	4
3277	3278	G6770	ECK1457	nhoA	X	N	Y	4
3282	3283	G6775	ECK1472	adhP	X	N*	Y	4
3289	3290	G6782	ECK1482	ddpX	X	N	Y	4
3293	3294	G6786	ECK1486	gadC	X	N	Y	4
3305	3306	G6798	ECK1504	lsrK/ydeV	X	N*	Y	4
3307	3308	G6800	ECK1506	lsrA/ego	X	N	Y	4

3308	3309	G6801	ECK1507	lsrC	X	N	Y	4
3309	3310	G6802	ECK1508	lsrD	X	N	Y	4
3310	3311	G6803	ECK1509	lsrB	X	N	Y	4
3311	3312	G6804	ECK1510	lsrF	X	N	Y	4
3312	3313	G6805	ECK1511	lsrG	X	N	Y	4
3317	3318	G6810	ECK1517	glsB/yneH	X	N	Y	4
3318	3319	G6811	ECK1518	sad/yneI	X	N	Y	4
3349	3350	G6842	ECK1579	speG	X	N	Y	4
3364	3365	G6857	ECK1594	mdtI	X	N	Y	4
3365	3366	G6858	ECK1595	mdtJ	X	N	Y	4
3366	3367	G6859	ECK1596	tqsA/ydgG	X	N	Y	4
3369	3370	G6862	ECK1601	folM	X	N	Y	4
3384	3385	G6877	ECK1630	dtpA/ydgR	X	N	Y	4
3386	3387	G6879	ECK1632	pdxY	X	N	Y	4
3393	3394	G6886	ECK1642	sodC	X	N	Y	4
3398	3399	G6891	ECK1647	gloA	X	N	Y	4
3413	3414	G6906	ECK1676	sufS	X	N	Y	4
3418	3419	G6912	ECK1683	menI/ydiI	X	N	Y	4
3424	3425	G6918	ECK1693	ydiO	X	N*	Y	4
3446	3447	G6940	ECK1742	astE	X	N	Y	4
3447	3448	G6941	ECK1743	astB	X	N	Y	4
3448	3449	G6942	ECK1744	astD	X	N	Y	4
3449	3450	G6943	ECK1745	astA	X	N	Y	4
3450	3451	G6944	ECK1746	astC	X	N	Y	4
3464	3465	G6958	ECK1769	ydjG	X	N	Y	4
3472	3473	G6967	ECK1779	yeaE	X	N	Y	4
3474	3475	G6969	ECK1781	yeaG	X	N	Y	4
3489	3490	G6984	ECK1796	leuE/yeaS	X	N	Y	4
3491	3492	G6986	ECK1798	dmlA/yeaU	X	N	Y	4
3504	3505	G6999	ECK1819	mntP/yebN	X	N	Y	4

3510	3511	G7005	ECK1831	msrC/yebR	X	N*	Y	4
3516	3517	G7011	ECK1837	pphA	X	N*	Y	4
3522	3523	G7017	ECK1858	znuA	X	N*	Y	4
3523	3524	G7018	ECK1859	znuC	X	N	Y	4
3539	3540	G7037	ECK1917	tcyL/yecS	X	N	Y	4
3540	3541	G7038	ECK1918	dcyD/yedO	X	N*	Y	4
3541	3542	G7039	ECK1919	tcyJ/fliY	X	N	Y	4
3556	3557	G7055	ECK1963	hchA	X	N	Y	4
3590	3591	G7091	ECK2023	ugd	X	N	Y	4
3597	3598	G7098	ECK2041	wcaJ	X	N	Y	4
3604	3605	G7105	ECK2054	wzc	X	N	Y	4
3605	3606	G7106	ECK2055	wzb	X	N	Y	4
3606	3607	G7107	ECK2056	wza	X	N	Y	4
3620	3621	G7123	ECK2082	yegS	X	N	Y	4
3622	3623	G7128	ECK2088	gatZ	X	N	Y	4
3623	3624	G7129	ECK2090	fbaB	X	N*	Y	4
3629	3630	G7135	ECK2096	thiD	X	N	Y	4
3631	3632	G7138	ECK2099	rcnA/yohM	X	N	Y	4
3634	3635	G7145	ECK2139	preT/yeiT	X	N	Y	4
3635	3636	G7146	ECK2168	lpxT/yeiU	X	N*	Y	4
3636	3637	G7147	ECK2169	mepS/spr	X	N	Y	4
3644	3645	G7158	ECK2238	yfaU	X	N	Y	4
3646	3647	G7160	ECK2240	rhmD/yfaW	X	N	Y	4
3650	3651	G7164	ECK2244	nudI/yfaO	X	N	Y	4
3652	3653	G7166	ECK2246	arnB/yfbE	X	N*	Y	4
3654	3655	G7168	ECK2248	arnA/yfbG	X	N	Y	4
3661	3662	G7175	ECK2262	rbn/elaC	X	N*	Y	4
3669	3670	G7184	ECK2284	alaA/yfbQ	X	N	Y	4
3670	3671	G7185	ECK2285	yfbR	X	N	Y	4
3672	3673	G7187	ECK2287	hxpA/yfbT	X	N	Y	4

3680	3681	G7195	ECK2297	folX	X	N	Y	4
3697	3698	G7212	ECK2335	fadJ/yfcX	X	N	Y	4
3698	3699	G7213	ECK2336	fadI/yfcY	X	N	Y	4
3715	3716	G7230	ECK2355	yfdR	X	N	Y	4
3719	3720	G7234	ECK2367	yfdE	X	N	Y	4
3721	3722	G7236	ECK2369	oxc	X	N	Y	4
3722	3723	G7237	ECK2370	frc	X	N	Y	4
3726	3727	G7241	ECK2374	lpxP/ddg	X	N*	Y	4
3727	3728	G7242	ECK2375	alaC/yfdZ	X	N	Y	4
3738	3739	G7254	ECK2387	mntH	X	N	Y	4
3743	3744	G7259	ECK2413	pdxK	X	N	Y	4
3747	3748	G7263	ECK2423	murQ/yfeU	X	N	Y	4
3750	3751	G7266	ECK2426	yfeX	X	N*	Y	4
3772	3773	G7288	ECK2453	eutD/eutI	X	N	Y	4
3777	3778	G7293	ECK2458	maeB	X	N	Y	4
3807	3808	G7325	ECK2527	iscS	X	N*	Y	4
3842	3843	G7365	ECK2626	rnlA/yfjN	X	N*	Y	4
3872	3873	G7399	ECK2664	alaE/ygaW	X	N*	Y	4
3878	3879	G7405	ECK2676	ygaZ	X	N	Y	4
3879	3880	G7406	ECK2677	ygaH	X	N*	Y	4
3882	3883	G7409	ECK2695	pncC/ygaD	X	N	Y	4
3883	3884	G7410	ECK2696	mltB	X	N	Y	4
3886	3887	G7415	ECK2729	pphB	X	N	Y	4
3893	3894	G7422	ECK2740	truD	X	N	Y	4
3894	3895	G7423	ECK2742	ispD	X	E	N	-
3918	3919	G7449	ECK2785	truC/yqcB	X	N	Y	4
3923	3924	G7454	ECK2806	csdA	X	N	Y	4
3925	3926	G7456	ECK2808	tcdA/ygdL	X	N	Y	4
3926	3927	G7457	ECK2809	mltA	X	N	Y	4
3953	3954	G7485	ECK2862	xdhA	X	N	Y	4

3954	3955	G7486	ECK2863	xdhB	X	N	Y	4
3955	3956	G7487	ECK2864	xdhC	X	N	Y	4
3969	3970	G7501	ECK2878	xanQ/ygfO	X	N	Y	4
3970	3971	G7502	ECK2879	guaD	X	N	Y	4
3971	3972	G7503	ECK2880	ghxQ/ygfQ	X	N*	Y	4
3974	3975	G7507	ECK2883	uacT/ygfU	X	N	Y	4
3975	3976	G7508	ECK2884	idi	X	N	Y	4
3983	3984	G7516	ECK2915	scpB/ygfG	X	N*	Y	4
3984	3985	G7517	ECK2916	scpC/ygfH	X	N	Y	4
3996	3997	G7530	ECK2949	rdgB/yggV	X	N	Y	4
3999	4000	G7533	ECK2958	mltC	X	N	Y	4
4007	4008	G7542	ECK2969	glcA/yghK	X	N	Y	4
4009	4010	G7544	ECK2973	glcE	X	N*	Y	4
4010	4011	G7545	ECK2974	glcD	X	N	Y	4
4023	4024	G7558	ECK2995	gpr/yghZ	X	N	Y	4
4026	4027	G7564	ECK3003	yqhD	X	N	Y	4
4027	4028	G7565	ECK3004	dkgA	X	N	Y	4
4039	4040	G7579	ECK3023	cpdA	X	N	Y	4
4048	4049	G7590	ECK3042	hldE/rfaE	X	N	Y	4
4052	4053	G7596	ECK3063	patA/ygjG	X	N*	Y	4
4078	4079	G7624	ECK3101	tdcG	X	N*	Y	4
4083	4084	G7631	ECK3120	kbaZ	X	N	Y	4
4093	4094	G7641	ECK3134	rsmI/yraL	X	E	N	-
4099	4100	G7647	ECK3141	yhbO	X	N	Y	4
4112	4113	G7662	ECK3186	kdsD	X	N	Y	4
4113	4114	G7663	ECK3187	kdsC	X	N	Y	4
4114	4115	G7664	ECK3188	lptC/yrbK	X	E	N	-
4115	4116	G7665	ECK3189	lptA/yhbN	X	E	N	-
4125	4126	G7676	ECK3211	nanK	X	N*	Y	4
4126	4127	G7677	ECK3212	nanE	X	N	Y	4

4162	4163	G7716	ECK3334	fkpA	X	N	Y	4
4168	4169	G7723	ECK3359	frlB	X	N	Y	4
4169	4170	G7724	ECK3360	frlC	X	N*	Y	4
4170	4171	G7726	ECK3361	frlD	X	N	Y	4
4186	4187	G7742	ECK3386	yrfG	X	N*	Y	4
4222	4223	G7826	ECK4117	dcuR	X	N	Y	4
4228	4229	G7833	ECK4135	yjeH	X	N	Y	4
4231	4232	G7836	ECK4140	epmB/yjeK	X	N	Y	4
4237	4238	G7842	ECK4158	orn	X	E*	N	-
4252	4253	G7858	ECK4192	ulaD	X	N	Y	4
4253	4254	G7859	ECK4193	ulaE	X	N	Y	4
4254	4255	G7860	ECK4194	ulaF	X	N	Y	4
4259	4260	G7865	ECK4203	fklB	X	N*	Y	4
4262	4263	G7868	ECK4207	qorB/ytfG	X	N	Y	4
4280	4281	G7888	ECK4254	lptF/yjgP	X	E	N	-
4281	4282	G7889	ECK4255	lptG/yjgQ	X	E*	N	-
4284	4285	G7892	ECK4259	idnO	X	N	Y	4
4285	4286	G7893	ECK4260	idnD	X	N	Y	4
4301	4302	G7910	ECK4286	yjhG	X	N	Y	4
4302	4303	G7911	ECK4287	yjhH	X	N*	Y	4
4311	4312	G7920	ECK4301	nanM/yjhT	X	N*	Y	4
4312	4313	G7921	ECK4302	nanC/yjhA	X	N*	Y	4
4337	4338	G7948	ECK4356	bglJ	X	N	Y	4
4344	4345	G7961	ECK1382	feaB	X	N*	Y	4
4346	4347	G812	ECK4232	nrdG	X	N	Y	4
4347	4348	G81	ECK3083	uxaC	X	N	Y	4
4355	4356	G8221	ECK3760	ilvG (pseudo)	X	N*	Y	4
4356	4357	G85	ECK2401	xapA	X	N	Y	4
4364	4365	M004	ECK3373	rpe	X	N	Y	4
4370	4371	M013	ECK0347	mhpD	X	N	Y	4

4371	4372	M014	ECK0348	mhpF	X	N	Y	4